Array based genotyping of medicinal plants from the genera *Salvia* and *Echinacea*

A thesis submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy

Alexandra Cristina Olarte Guasca

Master of Applied Microbiology and Biotechnology

School of Applied Science

RMIT University

December 2011

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Alexandra C. Olarte Guasca

12/2011

Acknowledgments

I wish to express my greatest gratitude to my supervisor Professor Eddie Pang for his exceptional mentoring and guidance throughout the years and to my second supervisor Dr. Gregory Nugent for his patience and guidance throughout this study.

I am also thankful to Dr. Hans Wohlmuth (Southern Cross University), to A/Prof. Reg Lehmann (MediHerb) and to Mr. Tim Groom (BRA) who were partners in this project.

I gratefully acknowledge the Australian Postgraduate Scholarship for offering me a scholarship and to RIRDC that funded this study.

I welcome the friendship and help of Dr. Nitin Mantri and Dr. Ruchira Jayasinghe who guide me in many of my experiments and cheered me up in difficult times.

A special thanks to Claudia Salazar and Viviana Ruiz for their assistance with the graphics in this study and for their constant encouragements and friendship.

Huge thanks to Ana Miranda and Dr. Jose Rodrigues for their help to print the thesis and for their unconditional friendship.

Finally, special thanks to my parents, my brother, Lucero and Raul whose confidence in me kept me going; and to my lab friends Debbie, Stephan, Eiseng, Carolina, Yuki, Urvi, Bijal and Hashmath for all their support and good times.

Thesis abstract

The medicinal herb industry has shown significant growth in developed countries over the past decade. For instance, Chinese medicinal herbs are becoming increasingly popular in Western countries for the treatment of modern diseases such as cardiovascular disease, asthma, and other long-term illnesses with minimum side effects. In addition, this growing interest in herbal medicines has also lead to the rediscovery of Amerind traditional medicines like *Echinacea*. Although the herbal medicines are gaining popularity, several reports on contamination, adulteration and misidentification have raised concerns on their commercialization. Therefore, an accurate identification of the herbs is crucial. DNA-based markers offer a possibility for an accurate identification (fingerprinting) of medicinal herbs by their genetic composition, which is not affected by agricultural and environmental factors.

This thesis documents the construction of a Subtracted Diversity Array (SDA) enriched with polymorphic and divergent DNA sequences for fingerprinting several economically important members of the genus *Salvia* and *Echinacea*. In order to construct the *Salvia*-specific SDA, suppression subtractive hybridization was performed between a pool of ten *Salvia* species and a pool of non-angiosperms and angiosperms (excluding the Lamiaceae) to selectively isolate *Salvia*-specific sequences. A total of 285 subtracted genomic DNA fragments were amplified and arrayed to construct this SDA. DNA fingerprints were obtained for fifteen *Salvia* genotypes including three that were not part of the original subtraction pool. Cluster analysis indicated that the *Salvia*-specific SDA was capable of differentiating closely related species of *S. officinalis* and

S. miltiorrhiza and was also able to reveal genetic relationships consistent with geographical origins. Species-specific features were also found for *S. elegans, S. officinalis, S. sclarea, S.* przewalskii, S. *runcinata* and *S. miltiorrhiza*. In addition, this SDA was able to fingerprint populations of *S. miltiorrhiza* and the genetic profiles obtained for each population significantly correlated with their chemical profiles obtained in previous studies.

Similarly, for the construction of the *Echinacea*-specific SDA, suppression subtractive hybridization was performed between a pool of twenty-four *Echinacea* lines and a pool of non-angiosperms and angiosperms (excluding the Asteraceae) to selectively isolate *Echinacea*-specific sequences that were amplified and arrayed. This SDA was capable of fingerprinting twenty-seven *Echinacea* lines including four that were not used in its construction. However, the use of this SDA for authentication purposes may be limited since only unknown samples of *E. paradoxa* and *E. purpurea* could be unambiguously identified in the cluster analysis. Furthermore, this *Echinacea*-specific SDA was also able to isolate highly polymorphic sequences some of which matched to known retrotransposons sequences. These sequences have the potential to become retrotransposon-based molecular markers useful for fingerprinting and studying diversity patterns in *Echinacea*.

The results from both SDA studies show that the enrichment of specific sequences during subtraction makes it possible to isolate a set of unique polymorphic sequences for the taxa under study which opens the possibility of fingerprinting species, populations and accessions that were not used to construct the array, with the advantage

that no prior knowledge of genetic sequence is needed. Therefore, SDA could be employed not only for authentication of species but also authentication within the same species. Furthermore, the results showed that SDA is a useful technique to find potential DNA markers for non-model plants that could assist not only for authentication but also for marker assisted selection of future plant breeding programs.

Thesis publications

Refereed Conference

Olarte, A., Mantri, N., Nugent, G., Li, C.G., Xue, C., Pang, E.C.K., 2010. Construction and validation of a prototype microarray for high-throughput genotyping of *Salvia* species. Plant & Animal Genome XVIII Conference, Town & Country Convention Center. San Diego, CA.

Olarte, A., Mantri, N., Nugent, G., Li, C.G., Xue, C., Pang, E.C.K., 2010. A gDNA microarray for genotyping Danshen (*Salvia Miltiorrhiza*) and other economically-important *Salvia* species. The 9th Meeting of the Consortium for Globalization of Chinese Medicine, Hong Kong on August 23-25, 2010.

Lau, E., Olarte, A., Mantri, N., Li, C.G., Xue, C., Pang, E.C.K., 2011. Fingerprinting provincial varieties of the Chinese red sage (*Salvia miltiorrhiza*) and the development of DNA-based markers for bioactive compounds. Plant & Animal Genome XIX Conference, Town & Country Convention Center. San Diego, CA.

Recently submitted manuscripts

Olarte, A., Mantri, N., Nugent, G., Wohlmuth, H., Li, C.G., Xue, C., Pang, E.C.K., 2011. A gDNA microarray for genotyping *Salvia* species. Submitted to Plant methods.

Mantri, N., Olarte, A., Li, C.G., Xue, C., Pang, E.C.K., 2012. Fingerprinting the Asterid Species using Subtracted Diversity Array reveals Novel Species-specific Sequences. Plos One.(7) 4.

Manuscript in preparation

Olarte, A., Lau, E., Mantri, N., Nugent, G., Sheng, S., Li, C.G., Xue, C., Pang, E.C.K., 2012. Fingerprinting of provincial varieties of the Chinese red sage (*Salvia miltiorrhiza*) and the identification of potential DNA-based markers for bioactive compounds. (Target date: 2012).

Olarte, A., Mantri, N., Nugent, G., Li, C.G., Xue, C., Pang, E.C.K., 2012. Fingerprinting of *Echinacea* species using the *Echinacea* Subtracted Diversity Array. (Target date: 2012).

Table of contents

List of tables

List of figures

List of abbreviations

µ	micro
^{0}C	degrees Celcius
AFLP	amplified fragment length polymorphism
APG	Angiosperm Phylogeny Group
BAC	bacterial artificial chromosome
bp	base pair
BSA	bovine serum albumin
CE	capillary electrophoresis
cDNA	complementary deoxyribonucleic acid
CoRAP	conserved region amplification polymorphism
CTAB	cetylmethylammonium bromide
Cy 3	cyanine-3
EST	Expressed Sequence Tags
DIG	digoxigenin
DArT	Diversity Array Technology
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
dUTP	deoxyuridine triphosphate
EDTA	ethylenediaminetetraacetic acid
EtBr	ethidium bromide
g	gram
GC	gas chromatography
gDNA	genomic DNA
h	hour
HCl	hydrochloric acid
HPLC	high performance liquid chromatography
ITS	internal transcribed spacer
ISSR	inter-simple sequence repeats

l	liter
LC/MS	liquid chromatography and mass spectrometry
LTR	long terminal repeat
m	meter
MS	mass spectrometry
*mat*K	gene believed to be involved in the splicing of group II introns
MgCl2	magnesium chloride
min	minute
ml	milliliter
mM	millimolar
mm	millimeter
MT	metric ton
MPLA	multiplexed ligation-dependent probe amplification
NaCl	sodium chloride
ng	nanogram
ORF	open reading frame
PLS	partial least squares
PCA	principal component analysis
PMT	photo multiplier tube
PCR	polymerase chain reaction
RFLP	restriction fragment length polymorphism
psaI	chloroplast gene for the photosynthetic system I-protein
*psb*A	chloroplast gene for the photosynthetic system II D1-protein
psbB	chloroplast gene for the photosynthetic system II 47 kDa protein
PVP	polyvinylpyrrolidone
RAPD	random amplified polymorphic DNA
*rbc*L	gene for the 1,5-bisphosphate carboxylase/oxygenase large subunit
rDNA	ribosomal DNA
rpm	revolutions per minute
*rpo*B	chloroplast gene coding for DNA-directed RNA polymerase beta.
SDS	sodium dodecyl sulphate
sec	seconds

SCAR	sequence characterized amplified region
SRAP	sequence related amplified polymorphism
SSR	simple sequence repeat
SNP	single nucleotide polymorphism
SPSS	Statistical Package for the Social Sciences
ss cDNA	single stranded cDNA
SSC	sodium chloride/sodium citrate buffer
SSPE	saline-sodium phosphate-EDTA buffer
STR	short tandem repeat
SDA	Subtracted Diversity Array
SSH	Subtraction Suppression Hybridization
SFE	supercritical fluid extraction
TLC	thin layer chromatography
TCM	Traditional Chinese Medicine
Tris	tri(hydroxymethyl) aminomethane
*trn*G, *trn*L	Gene coding for the transfer RNA (tRNA)

CHAPTER 1

Literature review

1.1 INTRODUCTION

The use of herbal medicines in developed countries is growing, for instance Chinese medicinal herbs are becoming increasingly popular worldwide. This increased number of people using herbal medicine products, have raised concerns on the safety of their use (Guo et al., 2009; Leung and Cheng, 2008; Zhao et al., 2006). Quality assurance is essential to ensure the safety and efficacy of herbal medicine. One of the main problems in quality assurance is the unequivocal identification of the herbs since contamination, adulteration, substitution or misidentification of the declared herbs is a common problem (Heubl, 2010; Joshi et al., 2004; Rotblatt, 1999; Sucher and Carles, 2008; Techen et al., 2004; Zhao et al., 2006). For example, *Salvia miltiorrhiza* Bge. which is one of the most popular medicinal herbs used in China is usually misidentified or adulterated with 17 different species of *Salvia* by local people in different regions of China (Li, 2008). Furthermore, herbal samples prepared from different parts of the plant, harvested from different geographical regions and seasons may result in products with variations in the concentration of bioactive components (Guo et al., 2009; Rotblatt, 1999). For instance, it has been found that different commercial products of *Echinacea*, a popular immunostimulant in Western countries, have considerable variation both between and within commercial batches (Wills et al., 2004). In order to understand the challenges involved to guarantee the identity of the herb and also the uniformity and

repeatability of their products, this review describes the different authentication techniques with their appropriate applications and limitations. A main emphasis is placed on *Salvia* and *Echinacea*, which are among the most popular herbs used in China and Western countries respectively.

1.2 CURRENT TRENDS IN MEDICINAL HERBS

The use of natural products for the prevention and treatment of various health problems has been in practice from ancient times (Sahoo et al., 2010). Traditional medicine is very popular in developing countries, for instance over 80% of the population in Asian and African countries depend directly on traditional medicine for primary health care (WorldHealthOrganization, 2008). In addition, there has been a growing interest in the therapeutical use of natural products. Medicinal herbal products are now widely sold on supermarkets and pharmacies, not only in specific health food stores, even insurance plans now cover herbal medicines (Rotblatt, 1999). Furthermore, herbal medicines are highly lucrative in the international market. Annual revenues in Western Europe reached US$5 billion in 2003-2004 and in China sales of products totaled US$14 billion in 2005 (WorldHealthOrganization, 2008).

The use of the herbal medicines in western societies was eclipsed by the development of organic chemistry and other medical advances such as antibiotics which resulted in a preference for synthetic products and increase in the economical power of the pharmaceutical companies (Baum et al., 2004; Rates, 2001). In contrast, traditional

medicines have been integral part in the health care and culture. For instance, Traditional Chinese Medicine (TCM) has an ancient history and has been highly treasured as a precious cultural heritage (Li, 2008). National surveys performed in the past 30 years indicated that a total of 12,807 medicinal species are used in China (Leung and Cheng, 2008). TCM has been used to treat modern diseases such as cardiovascular disease, asthma, and other long-term illnesses (Zhang et al., 2007). As an example, the roots and rhizome of *Salvia miltiorrhiza* Bge, known as danshen, is officially listed in the Chinese Pharmacopeia to be used for the treatment of menstrual disorders, menorrhagia, insomnia, menostasis, blood circulation diseases and other cardiovascular diseases (Cai et al., 2002). The therapeutic effects and minimum side effects of many herbal remedies have increased the interest in the study of traditional Chinese medicines. As well, many multinational pharmaceutical companies are developing Chinese medicine with Chinese companies (Zhang et al., 2007)**.** This growing interest in herbal medicines has also lead to the rediscovery of western traditional medicines. *Echinacea* for example, which is native to the Canadian prairies and the prairie states of the United States, has a long tradition as a folk medicine for the Native Americans. However before 1980, almost all laboratory and clinical evaluations were conducted in Germany. Nowadays, *Echinacea* are among the top 10 selling herbal medicines in the U.S. and Europe (Yu and Kaarlas, 2004).

Although the herbal medicines are gaining popularity, several reports on contamination (heavy metals/toxic components), adulteration/substitution (deliberate addition of synthetic compound or substitution with similar herbs), and misidentification or mislabeling (similarity in appearance, confused nomenclature) have raised concerns on

their commercialization (Drasar and Moravcova, 2004; Rotblatt, 1999; Techen et al., 2004; Zhao et al., 2006). Therefore, authentication of medicinal plants is critical for protection of public and industry. The medicinal importance of the genus *Salvia* and *Echinacea* will be discussed in the following sections together with the main misidentification and adulteration problems found in these two genera.

1.2.1 Commercial and medicinal value of the genus *Salvia*

Salvia (Lamiaceae) is an important genus with approximately 1000 species distributed widely in many regions of the world including the Mediterranean area, southern Africa, Central and South America, and Asia (Walker et al., 2004). The genus exhibits a wide range of morphological (diversity of staminal structure and floral morphology) and ecological variation which has made taxonomical classification difficult (Bruna et al., 2006). One of the most widely accepted classifications, Bentham's (1848), separated the genus into 12 sections; however 500 new species has been recognized since this study was performed. Recent molecular phylogenetic studies have recognized three major lineages; each related to other genera of the tribe Menthae (Walker et al., 2004).

This genus is largely cultivated for ornamental, culinary and medicinal uses (Bruna et al., 2006; Topcu, 2006). For example, *S. officinalis* L. is used to preserve foods and employed as a spice for flavouring (Topcu, 2006). Its essential oil is used in perfumery and cosmetics (Echeverrigaray and Agostini, 2006). Furthermore, the plant and their derivates are known to have a wide range of biological activities, such as antibacterial, antioxidative, anti-inflammatory, hypoglycemic activity (potential anti-diabetic) and cholinesterase inhibitory which is relevant to the treatment of Alzheimer disease.

Another example is *Salvia fruticosa* Mill. which it is known from their medicinal properties since ancient times by Greek, Spanish and Morrocan Arab herbalists. The infusion of leaves and shoots are used for relieving headaches, rheumatic pains, stomach ache, colds and cough; it also has the reputation of being blood depurative, antiseptic and sedative (Rivera et al., 1994). **Table 1.1** summarizes some of the most popular species used for their medicinal purposes. It is important to note that these species were also used during the development of this study.

Although the *Salvia* species have important medicinal and commercial value, a correct identification of the species is challenging. For example, danshen is one of the most popular medicinal herbs in China with a market value of about 2 billion of US dollars in 2005 (Yu et al., 2007). However, field investigations indicated that roots of 17 different species of *Salvia* were used as danshen by local people in different regions of China even though that the Chinese Pharmacopoeia describes danshen as the dried root and rhizome of *Salvia miltiorrhiza* **(Figure 1.1)** (Li, 2008). Even some of these closely related species sold as danshen are known to have other therapeutic effects, for example *S. przewalskii* Maxim. is used to remove heat from the blood and relieving swelling **(Table 1.1)**. Another example of misidentification is found among *S. fruticosa* and *S. officinalis*, as discussed above both species have an important medicinal value, however their morphological similarity and the fact that hybrids between them can occur has made it difficult to find characters that allow an accurate identification. To date, the type of calyx hair has been found to be the most useful character to distinguish between this two species and its hybrids (Dudai et al., 1999; Reales et al., 2004). Therefore, in order to avoid misidentification among closely related species of *Salvia* the use of

authentication techniques is necessary, since many of the related species may have different bioactive components and thus their pharmacological activity could differ from the correct herb.

Table 1.1. List of some of the most popular *Salvia* species used for their medicinal purposes.

Species	Native to	Uses	Reference
S. elegans	Central America	Balance the nervous system	(Mora et al., 2006)
		Potential antidepressant and anxiolytic activity	(Herrera-Ruiz et al., 2006)
S. fruticosa	Eastern Mediterranean, North Africa and western Asia	Infusions used to lower blood pressure and blood sugar levels.	(Topcu, 2006)
		Bacteriostatic and bactericidal activities	(Longaray-Delamare et al., 2007)
S. lavandulifolia	Spain and southern France	Hypoglycemic activity	(Zarzuelo et al., 1990)
S. mexicana	Central Mexico	Antioxidant and anti-inflammatory activities	(Topcu, 2006)
S. miltiorrhiza	China	Could improve microcirculation, dilate the coronary arteries, increase blood flow and prevent myocardial ischemia.	(Zhou et al., 2005)
		Potential benefits for lipid control	
		Used in the treatment of acute stroke.	(Yu et al., 2007)

S. officinalis	Mediterranean Region	Relieve of inflammation of the oral cavity and throat	(Topcu, 2006)
		Bacteriostatic and bactericidal activities	(Longaray-Delamare et al., 2007)
		Hypoglycemic activity	(Alarcon-Aguilar et al., 2002)
		Studied for the treatment of Alzhermer's disease	(Savelev et al., 2004)
S. przewalskii	China	Used by locals as Qinjiao, which has the effects of expelling wind, dredging and activating the channels and collaterals, removing the heat from the blood and relieving swelling	(Li, 2008)
S. runcinata	South Africa	Excential oil presented anti-inflammatory and anti-malarial activity	(Kamatou et al., 2008)
S. sclarea	Northern Mediterranean, North Africa and Central Asia	Used to treat symptoms associated to menopause and the essential oil used in the cosmetic industry.	(Topcu, 2006)
		Extracted diterpentoids have shown effective antimicrobial activity (potential anti-biofilm agent)	(Kuzma et al., 2007)

Figure 1.1. Image of *Salvia miltiorrhiza* Bunge. Left: plants; right top: leave; right middle: inflorescence; and right bottom: roots (Sheng, 2007).

1.2.2 Medicinal and commercial value of the genus *Echinacea*

The genus *Echinacea* native to the Canadian prairies and the prairie states of the United States is a herbaceous, perennial flowering plant **(Figure 1.2)** (Mazza and Cottrell, 1999). The plant was used by the Plain Indians, particularly *E. angustifolia* DC., for relieving toothache, coughs, colds, sore throats, snakebites and as a painkiller (Kindscher, 1989). Studies on this genus have shown that it could effectively moderate the incidence, duration and severity of symptoms associated with common cold (Percival, 2000; Yu and Kaarlas, 2004). In addition, *Echinacea* is well known for its ability to stimulate the immune system; it has been proven to be able stimulate various immune cells (Barrett, 2003). Recent studies are focusing on understanding how this

immunostimulation can lead to enhance resistance to infectious diseases (Altamirano-Dimas et al., 2007).

E. angustifolia, *E. purpurea* (L.) Moench. and *E. pallida* (Nutt.) Nutt., are the three main species commonly used for extracts or whole-plant products in the natural medicine industry. For example, in Australia where is reported that 50% of the population uses complementary and alternative medicine the annual consumption of *E. purpurea* in 2000 was 80MT, 15 to 20MT of *E. angustifolia* and 1MT of *E. pallida*. In the U.S. alone, an annual sales of *Echinacea* products has been estimated to be from more than $200 to $300 million (Barrett, 2003; Yu and Kaarlas, 2004). However, this increase in the market led to inadequate quality control. For instance, many *Echinacea* samples were found to be adulterated with *Parthenium integrifolium* L. roots, since its roots are larger and easier to harvest (Yu and Kaarlas, 2004). This problem is not a major concern anymore due to the commercial cultivation of these three species in places such as North America, Europe and Australia (Chuang et al., 2010a). However, other *Echinacea* species may appear in medicinal products due to the introduction of wild collected seeds into cultivation without proper authentication (Binns et al., 2002b). Therefore, species misidentification remains problematic since there are many morphological similarities between species, in particular between *E. pallida* and *E. angustifolia*. For instance, in Europe *E. pallida* was sold as *E. angustifolia*. This misidentification was due to the high morphological variability find within populations which made difficult the use of the identification keys proposed by McGregor's taxonomic classification (Binns et al., 2002a).

McGregor's classification (McGregor, 1968) was based on morphological traits and chromosome numbers performed on a wide-scale sampling of wild populations which he grew in an experimental garden and greenhouse over a time period of eight years. During his work he recognized nine species **(Table 1.2)**, including two varieties of *E. angustifolia* (*E. angustifolia* DC. var. *angustifolia* and *E. angustifolia* DC. var. *strigosa* McGregor) and two of *E. paradoxa* (Norton) Britton (*E. paradoxa* (Norton) Britton var. *paradoxa* and *E. paradoxa* (Norton) Britton var. *neglecta* McGregor). Additionally, he found that all taxa could hybridize when brought together and also found natural hybrids, which implies a high level of gene flow.

Figure 1.2.
Image of *Echinacea purpurea*

Table 1.2. Taxonomy of McGregor compared to the revised taxonomy of Binns et al, 2002 for species and varieties of genus *Echinacea.*

Mc Gregor (1968)	***Binns et al.(2002)***
E. purpurea (L.) Moench.	*E. purpurea* (L.) Moench.
E. pallida (Nutt.) Nutt.	*E. pallida* var. *pallida* (Nutt.) Nutt. var. *pallida*
E. angustifolia DC.	*E. pallida* (Nutt.) Nutt. var. *angustifolia* (DC.) Cronq
E. sanguinea Nutt.	*E. pallida* Nutt.) Nutt. var. *sanguinea* (Nutt.) Gandhi & R.D. Thomas
E. simulata McGregor.	*E. pallida* (Nutt.) Nutt. var. *simulata* (McGregor) Binns, B.R. Baum, & Arnason
E. tennesseensis(Beadle) Small.	*E. pallida* var. *tennesseensis* (Beadle) Binns B.R.Baum, & Arnason
E. atrorubens Nutt.	*E. atrorubens* (Nutt.) Nutt. var. *atrorubens*
E. paradoxa (Norton) Britton var. *paradoxa*	*E. atrorubens* (Nutt.) Nutt. var. *paradoxa* (J.B. Norton) Cronq.
E. paradoxa (Norton) Britton var. *neglecta* McGregor	*E. atrorubens* (Nutt.) Nutt. var. *neglecta* (McGregor) Binns, B.R. Baum, & Arnason
E. laevigata (Boynton & Beadle) Blake	*E. laevigata* (C.L. Boynton & Beadle). S.F. Blake

However, this classification have inconsistencies among the descriptions and practical difficulties using his keys have been found (Binns et al., 2002a). A revision of this classification performed by Binns et al. (2002a) was based on the measurements of morphometric traits using 321 individuals which were sampled in the wild and grown in a green house. The statistical analysis supported two possible cluster solutions one in which *E. purpurea* is the sole taxon in a subgenus called *Echinacea* and all the other taxa are included in a subgenus called *Pallida*. The second cluster solution supported a classification of four species and eight varieties **(Table 1.2)**. The revised taxonomy

(Binns et al., 2002a) recognized all of McGregor's taxa except for the variety *E. angustifolia* DC. var. *strigosa* McGregor. However, results from other investigations do not completely support neither of these classifications (Kim et al., 2004; Mechanda et al., 2004a; Wu et al., 2009). To date, for commercial purposes the classification sensu McGregor still being employed until other studies are able to give better support to current or new classifications, since any re-labeling of the *Echinacea* products will bring cost to the industry and will produce confusion to the customers (Blumenthal and Urbatsch, 2006).

1.3 FACTORS AFFECTING THE QUALITY OF HERBAL PLANTS

Unequivocal species authentication is a critical step in quality assurance, since it will ensure that the right plant material will be used for a specific treatment. However, the correct species identification of the herb cannot always ensure the level of the phytochemicals in the plant material (Baum et al., 2001). Standardization of active components on herbal medicine is a challenging task since is difficult to control the factors that could affect the chemical composition. For example environmental conditions (sunlight, temperature, soil), developmental variations (age of plant, harvest stage) and manufacturing process (drying, extracting, storing) can affect the concentration of active ingredients of the product (Rotblatt, 1999). For instance, three bioactive components of *Echinacea*, cichoric acid and dodeca 2E, 4E, 8Z, 10E/Z-tetraenoic acid isobutylamide (alkamides 8, 9), were quantified in 25 *Echinacea*-containing remedies. The quantity of the active components varied depending on the remedy, the *Echinacea* species, the part of the plant used, it even varied between

different batches of the same remedy (Osowski et al., 2000). Other studies also support these findings, for instance, Binns et al. (2002b) studied the phytochemical variation from roots and inflorescences in wild and cultivated populations in all the species of *Echinacea*, the highest amounts of cichoric acid were found in older, wild inflorescences of *E. sanguinea* Nutt. (*E. pallida* var. *sanguinea*).

This phytochemical variation has also been found in *Salvia*. For example, it has been found that the bioactive components of the roots of *S. miltiorrhiza* change in relation to harvest time and the geographical population/germplasm line (He et al., 2010; Li et al., 2009a; Sheng et al., 2009). The results obtained by Li et al. (2009) showed that there are significant differences in the content of biomarker compounds across different geographical populations even when they were cultivated under the same conditions and harvested at the same time, which imply that the selection of a specific Danshen-population with good agronomical and phytochemical characteristics is an important factor if the levels of phytochemical content need to be standardized. Therefore, in order to improve the quality of the medicinal herbs optimization of the germplasm line is as important as optimizing the cultivation area, harvest stage, post harvest and manufacturing handling.

1.4 AUTHENTICATION TECHNIQUES

Authentication techniques can help to avoid contamination, substitution and misidentification of the plant material starting form the collection of the raw material up to the finished product (Techen et al., 2004). These techniques are also fundamental to

evaluate and standardize the levels of phytochemicals in the herb (Zhao et al., 2006). As explained in the above sections, adulteration/substitution, misidentification and phytochemical variation are a major concern in the commercialization of *Salvia* and *Echinacea* species. On the following sections each of authentication techniques will be explained focusing on the applications for these two genera. **Figure 1.3** summarizes the main advantages and disadvantages of each of these techniques.

1.4.1 Macroscopic and microscopic identification

Macroscopic identification is performed based on parameters like shape, size, color, texture, odor and taste that are compared to a standard reference material (Joshi et al., 2004)**.** Its main advantage is that is a simple and fast authentication technique. The disadvantages of this method are that it may not allow the identification of closely related species or varieties and it requires a skilled person and access to herbaria references (Techen et al., 2004; Zhao et al., 2006).

Microscopy uses comparative inspection of whole, sliced and powdered material. The examination focuses on unique hairs (e.g. glandular or stellate), cell types, fibers, granular objects and minute floral and fruit characteristics of the plant material (Techen et al., 2004). Apart from being useful for identifying processed herbs, this technique is also useful for distinguishing species with similar morphological features (Zhao et al., 2006). For instance, the presence of short (0.05-0.6 mm) adpressed eglandular hairs could be used to distinguish *S. officinalis* from other closely related species such as *S. fruticosa* and *S. blancoana* (Reales et al., 2004). However this technique is not useful in

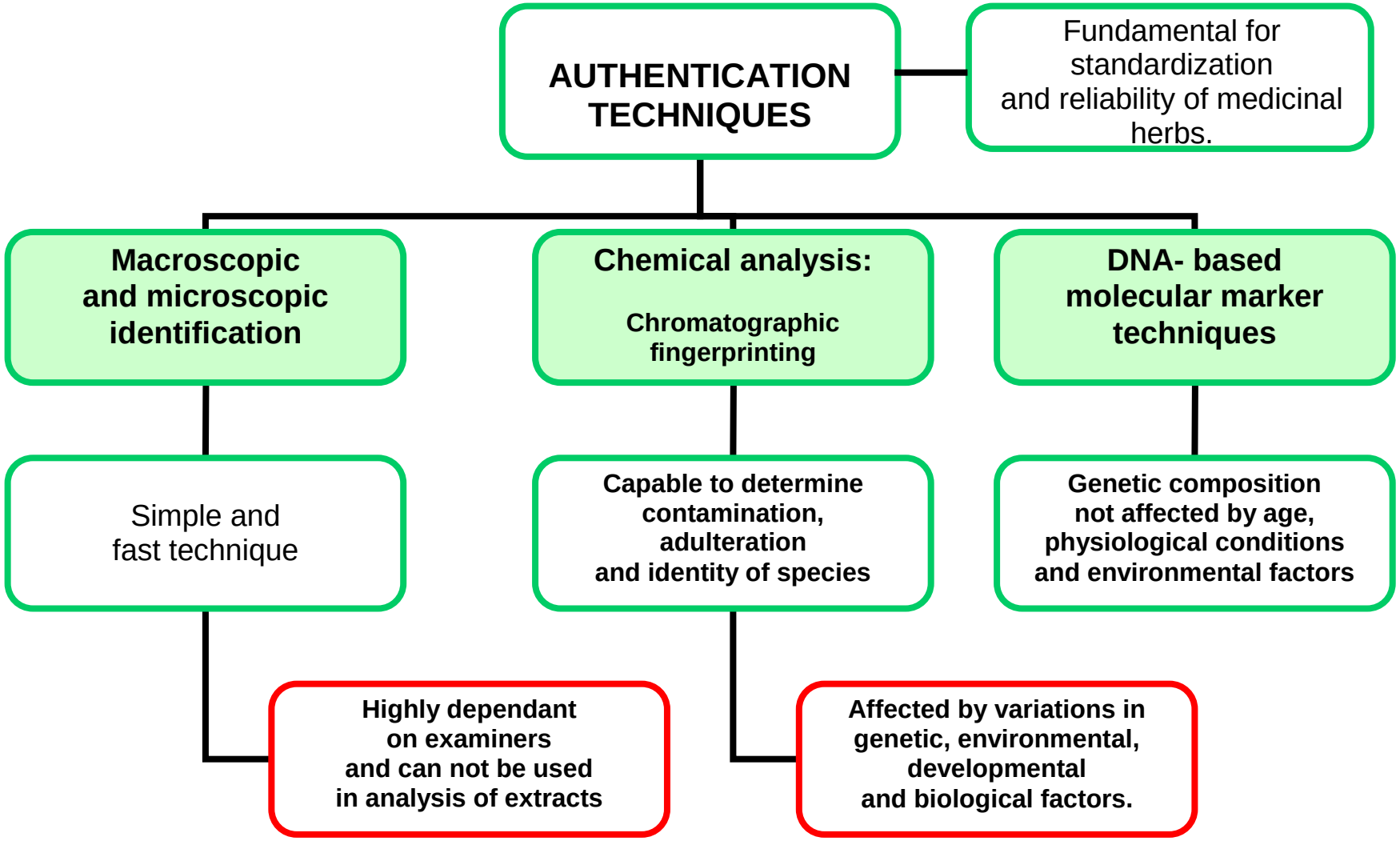


Figure 1.3. Advantages and disadvantages of authentication techniques.

the analysis of extracts, since all the identifiable characteristics will be lost during extraction (Zhao et al., 2006).

1.4.2. Chemical analysis

Thin layer chromatography (TLC), gas chromatography (GC), high performance liquid chromatography (HPLC) and capillary electrophoresis (CE) are among the most commonly analytical method used for chemical authentication (Fan et al., 2006; Obradovic et al., 2007). These methods achieve accurate plant identification by obtaining a chromatographic fingerprinting that is based on the presence or absence of characteristic peaks specific for the herb being analyzed, ensuring that related herbs or known adulterants will have different fingerprints (Zhao et al., 2006). In addition, these chromatographic separations could be coupled with Mass Spectrometry (MS). For instance liquid chromatography and mass spectrometry are used to characterize active components such as peptides/protein, carbohydrates and nucleic acids (LC/MS). Furthermore, MS can be also used to detect residual contaminants in the herbs such as steroids, pesticides, heavy metals and illegally-added synthetic drugs (Cai et al., 2002).

1.4.2.1 *Salvia*

The major secondary metabolites in *Salvia* are terpenoids, phenolic acids and flavonoids. The aerial parts contain mainly flavonoids, triterpenoids and monoterpenes while the roots contain mainly diterpentoids (Topcu, 2006). Among hydrophilic components in *Salvia* the phenolic acids, mainly caffeic acids, constitute a major part. Many of these phenolic acids have been identified and found unique from the Chinese

Salvias, *S. miltiorrhiza*, *S. chinensis* and *S. yunnanensis*; consequently they were designated salvianolic acids A-K and yunnaneic acids A-H in order to reflect the origin (Lu and Foo, 2002).

HPLC is a popular method for analysis of secondary metabolites in *Salvia*. For instance, it has been extensively used for the quality control of the roots of *S. miltorrhiza* (danshen) and derived pharmaceutical preparations by detecting lipophilic and hydrophilic active components (Liu et al., 2007; Ma et al., 2007; Yang et al., 2006a). The diterpenoid tanshinones are the major constituents of the lipid fraction in danshen and are recognized as the principal bioactive constituent. More than 30 diterpenoid tanshinones have been isolated of which three, Cryptotanshinone, tanshinone I and tanshinone IIA, are the most studied and used for quality control. The major active ingredients in the water soluble fraction include phenolic acids of which salvianolic acid B, danshensu (salvianolic acid A) and photocatechuic aldehyde constitute the most abundant (Cai et al., 2002; Hu et al., 2005; Wang and Wu, 2010; Zhou et al., 2006). According to the Chinese Pharmacopeia, tanshinone IIA and salvianolic acid B have been selected as diagnostic marker compounds for danshen. These two compounds in dry danshen should not be less than 2 mg/g and 30mg/g respectively. HPLC analyses have been able to determine that *S. miltiorrhiza* is the only species which roots are able to meet the requirements of Chinese Pharmacopeia (Li, 2008). Therefore, HPLC is capable to differentiate the roots of *S. miltiorrhiza* from roots of closely related species that are commonly misidentified as danshen (Li, 2008; Zhong et al., 2009).

1.4.2.1 Echinacea

A wide variety of secondary metabolites have been identified in *Echinacea*. The components that have received more attention and that have been associated with most biological activities are alkamides, polysaccharides, glycoproteins, polyacetylenes, and caffeic acid derivates (Arnason et al., 2002). As for the *Echinacea* volatile components terpenoids predominate the aerial parts while the root tissues are rich in aldehydes, terpenoids, miscellaneous compounds, and alcohols (Mazza and Cottrell, 1999).

Different metabolic compounds have been used as unique qualitative markers for species identification of root materials in industry. For example, cichoric acid has been used as a marker for *E. purpurea* and echinacoside as a marker for *E. angustifolia* and *E. pallida* (Arnason et al., 2002). Also ketoalkene/ynes have been used as markers for *E. pallida* (Baum et al., 2006). However, the presence/absence of these markers is not the best choice for species identification, for instance trace amounts of echinacoside were found in *E. purpurea* (Baum et al., 2006). Therefore, the species differentiation is presently performed on the basis of distribution and relative content of metabolites. For example, differentiation of the three most commonly used *Echinacea* species (*E. purpurea*, *E. angustifolia* and *E. pallida*) has been achieved by different techniques such as quantitative HLPC (Bergeron et al., 2000), supercritical fluid extraction (SFE) coupled with GC/MS (Hou et al., 2010) and reverse face- HPLC (Laasonen et al., 2002). Even spectroscopy analyses such as H Nuclear magnetic resonance spectrometry have been used (Frederich et al., 2009).

Analysis of the metabolic profiling of all *Echinacea* species has also been performed. For instance, *Echinacea* roots and flowerheads samples of wild and cultivated populations were analyzed using reverse-phase HPLC in order to quantify alkamides, ketoalkenes/ynes and phenolics in the populations that represented all *Echinacea* species (Binns et al., 2002b). This phytochemical profile obtained was used to classify the populations, which clustered accordingly to Binns et al. (2002a) taxonomical classification. However, another study performed by Wu et al. (2009) obtained a metabolic profiling with 43 lipophilic metabolites which classified 40 geographically and morphologically diverse *Echinacea* populations in agreement with McGregor' classification (Wu et al., 2009). Although both studies were able to distinguish the species and varieties based on their chromatographic profiles, additional populations will have to be analyzed in order to validate their results.

Chromatographic fingerprinting has the potential to determine contamination, adulteration and identity of species in *Salvia* and *Echinacea*. However, these techniques are dependent on the proper identification of compounds unique for the plant material, which can vary depending on the source of the herb, environmental, developmental and biological factors as well as processing methods (Cai et al., 2002; Drasar and Moravcova, 2004). For instance, hybrids between *S. officinalis* and *S. fruticosa* may occur spontaneously or from breeding programs. A hybrid obtained by crossing high quality *S. officinalis* clones with *S. fruticosa* has an intermediate morphology between the two parents but the essential oil resembles *S. officinalis* (Dudai et al., 1999). Therefore, accurate plant identification could not be achieved the by chemical composition of the essential oil between *S. officinalis* and the hybrid. In addition, many

of the secondary compounds are extracted from plant parts that could take months to develop; such is the case of the roots of *S. miltiorrhiza* which are harvested after 150-290 days after planting (Sheng et al., 2009). This is a great disadvantage for plant breeding programs which need faster screening techniques.

1.4.3 DNA-based molecular marker techniques

Molecular markers generally refer to biochemical constituents, including primary and secondary metabolites and other molecules such as proteins and nucleic acids (Kiran et al., 2010). The DNA markers, which are the more commonly used, are considered as more reliable tools for identification of species than chromatographic methods since the genetic composition of the plant is not affected by age, physiological conditions or environmental or developmental factors (Joshi et al., 2004; Kiran et al., 2010). However, DNA-based analyses of medicinal herbs have raised several challenges since they are not model plants, their genome sizes are usually unknown and there is a lack of molecular-marker based approaches for medicinal plant improvement (Canter et al., 2005). Therefore, the ideal molecular marker for authentication of medicinal herbs should fulfill the following criteria:

- **Polymorphic among the taxa:** The marker should be able to differentiate among closely related species or varieties, and common botanical adulterants in order to ensure the source of the plant material (Baum et al., 2001; Joshi et al., 2004).
- No require prior knowledge of **sequence information:** which will facilitate the development of the method on any species.

- **Associated or linked to agricultural traits:** The marker could be used for the prediction of desirable traits such as high concentration of bioactive components at an early stage of the reproductive cycle, allowing the selection of optimum genotypes. Compared to conventional breeding, the use of **marker-assisted selection** can reduce development cost and time required for evaluation of crosses (Canter et al., 2005)**.**

The DNA marker analysis can be classified on three classes based on the method of their detection: polymerase chain reaction (PCR)-based methods, sequencing-based methods and hybridization–based methods (Joshi et al., 2004).

1.4.3.1 PCR-based methods

These methods involve amplification of particular DNA sequences or loci by PCR; such amplifications can generate a genetic fingerprint for a given plant or specific species. A fingerprint with a high number of polymorphic amplifications is more useful since they will give the uniqueness to the fingerprint that will allow an accurate identification of the plant samples analyzed (Joshi et al., 2004; Techen et al., 2004). As it can be seen in **Table 1.3**, different types of PCR-based methods such random amplified polymorphic DNA (**RAPD**), amplified fragment length polymorphism (**AFLP),** simple sequence repeat polymorphism (**SSR**), inter-simple sequence repeats (**ISSR**) and sequence characterized amplified region (**SCAR**) have been employed for genetic diversity studies on different species, populations and lines of *Salvia* and *Echinacea*. However, none of these markers is ideal; each marker has its own advantages and disadvantages as explained in **Table 1.3**. For example SSR, ISSR and SCAR are sequence-specific

molecular markers meaning they need previous knowledge of sequence in order to design the specific primers; instead markers such as RAPD and AFLP can overcome this limitation by the use of arbitrary primers (Agarwal et al., 2008). **Table 1.3** shows that RAPD and AFLP markers have been more extensively used in the study of these two genera and that RAPD markers have also been used as fast and reliable method for authentication of the three commercially relevant *Echinacea* species *E. purpurea*, *E. angustifolia* and *E. pallida* (Wolf et al., 1999). Moreover, RAPD and AFLP markers have been found associated to the content of bioactive components. For instance, relationships between essential oil content and RAPD genotyping profiles were found in *S. officinalis* (Böszörményi et al., 2009) and *S. fruticosa* (Skoula et al., 1999). Also, RAPD markers were able to predict polyphenol content in aerial parts of *E. purpurea* (Chen et al., 2009a) and AFLP DNA fingerprints were found to be statistically significant as predictors of cichoric acid and alkamides 8 and 9 in cultivated *E. purpurea* and some related wild species (Baum et al., 2001). Although PCR–based methods have several advantages for authentication of medicinal herbs (particularly RAPD and AFLP markers) they also have important disadvantages. For example, most of these methods are based on gel electrophoresis, which is time consuming, and correlating bands on gels for the allelic variants present great difficulty that can lead to inaccurate interpretations (Jaccoud et al., 2001). In addition, there is not guarantee of genetic identity between two bands of the same size unless they are sequenced or analyzed by Southern blots (Kingsley et al., 2002).

1.4.3.2 Sequencing-based methods

This approach has been used for species identification by comparing sequences of the same locus in order to find variations (insertion, deletion, inversion) on the DNA sequence which allows the discrimination among individuals (Joshi et al., 2004). The use of this method for authentication of medicinal plants is also found in literature in the context of phylogenetic studies and as a general effort that aims to barcode all land plants (Sucher and Carles, 2008). "DNA barcoding" aims to provide rapid, accurate and automatable species identification by sequencing standardized gene regions that have been amplified from DNA extracted from an unknown specimen and compared this sequence to a reference sequence library from known species (Bruni et al., 2010; Hebert and Gregory, 2005). Most of the reports have analyzed sequences obtained from the nuclear and chloroplast regions. Among the nuclear region the variable internal transcribed spacer (ITS) ribosomal DNA (rDNA) and 5S rDNA gene are most commonly used for authentication of medicinal herbs (Chiou et al., 2007; Gnavi et al., 2009; Sucher and Carles, 2008). In contrast, chloroplast regions such as *rbc*L, *mat*K, *rpo*B and intergenic spacers between *trn*H-*psb*A and *trn*L-*trn*F are more popular for phylogenetic and barcoding analyses (Fazekas et al., 2008; Sucher and Carles, 2008).

Previous studies have found genetic variation among *Salvia* species using sequences from ribosomal and chloroplast regions. For instance, ITS regions have been able to differentiate *S. miltiorrhiza* from other non-danshen species and were also found to differentiate among *S. miltiorrhiza* populations (Xu et al., 2009). Furthermore, intergenic spacers between *trn*H-*psb*A and *trn*L-*trn*F together with *rbc*L and the nuclear ITS region have been used to conduct molecular phylogenetic analyses in the tribe

Table 1.3. PCR-based molecular marker techniques employed for genotyping *Salvia* and *Echinacea.*

Molecular marker	Explanation	Applications on *Salvia* and *Echinacea*
Random amplified polymorphic DNA (**RAPD**)	This technique is based on one primer with an arbitrary sequence of about 10 base pairs that is able to bind in multiple sites within the genome(Techen et al., 2004). **Advantage:** Easy to develop. **Disadvantage:** Reproducibility is very low.	***Salvia***: Used to evaluate the genetic diversity of *Salvia hispanica* L.(Cahill, 2004), *Salvia fruticosa* Mill. (Skoula et al., 1999) and *S. officinalis* (Böszörményi et al., 2009; Echeverrigaray and Agostini, 2006). Also used to characterize several African and American sage species (Bruna et al., 2006). ***Echinacea***: RAPD have been used to study the genetic relationship and diversity of the three commercially relevant *Echinacea* species (Kapteyn et al., 2002). Also have been proposed as a rapid and reliable test for proving the identity of these three species (Wolf et al., 1999).
Amplified fragment length polymorphism (**AFLP)**	Reliable technique, which includes digestion of total DNA, ligation of adaptors at the end of the restriction fragments, and selective PCR amplification using primers that bind to the adaptors (Sucher and Carles, 2008). **Advantage:** Detects thousands of independent loci. **Disadvantage:** Expensive and time consuming if applied in big populations (Joshi et al., 2004; Techen et al., 2004).	***Salvia***: Employed to perform a genetic diversity analysis on twenty seven *S. miltiorrhiza* geographical populations (Wang et al., 2007). ***Echinacea***: Have been used to asses the genetic diversity and phenetic relationships of natural populations and commercial lines of *Echinacea* (Mechanda et al., 2004a). Also employed to detect species-specific fragments for each of the commercially relevant *Echinacea* (Russi et al., 2009).

Molecular marker	Explanation	Applications on *Salvia* and *Echinacea*
Simple sequence repeat polymorphism (**SSR**) or micro satellites	SSR are short-sequence motifs consisting of 2 or more nucleotides, which repeat in tandem (Sucher and Carles, 2008). Due to the hypervariability of the repeats, primers specific to the regions flanking the microsatellites are expected to generate amplicons which will vary in length among individuals. **Advantage:** Robust and reliable **Disadvantage:** Needs previous sequence knowledge (Agarwal et al., 2008).	***Salvia***: SSR derived from *S. miltiorrhiza* EST sequences were able to detect genetic diversity among test samples of *S. miltiorrhiza* and distinguish it from other *Salvia* plants (Deng et al., 2009). Also, SSR markers have been developed from EST sequences of *S. fruticosa* which have been used to study genetic structure of *S. officinalis* (Mader et al., 2010). ***Echinacea***: At the time of the writing a search in the public literature did not find any study using SSR on *Echinacea.*
Inter simple sequence repeats (**ISSR**)	ISSR is a specific primer-based polymorphism detection system where primers anchored at a particular SSR is used to amplify the DNA between two flanking SSR (Joshi et al., 2004; Sucher and Carles, 2008). **Advantage:** More reproducible than RAPDs and less expensive than AFLPs (Techen et al., 2004). **Disadvantage:** Previous knowledge of sequence.	***Salvia***: ISSR were used to asses the genetic diversity of five cultivated populations of *S. miltiorrhiza*. Five ISSR primers amplified a total of 120 bands, all of them were found polymorphic (Song et al., 2010). ***Echinacea***: At the time of the writing a search in the public literature did not find any study using ISSR on *Echinacea.*
Sequence characterized amplified region (**SCAR**)	This technique uses the polymorphic amplicons obtained by RAPD. Then the sequence of a unique amplicon is used to design specific primers (Kiran et al., 2010). **Disadvantage:** Needs previous sequence knowledge.	***Salvia***: At the time of the writing a search in the public literature did not find any study using SCAR on *Salvia.* ***Echinacea***: A SCAR marker unique for *E. purpurea* was developed from a RAPD marker. The marker did not amplify for *E. pallida* or *E. angustifolia* (Adinolfi et al., 2007)

Mentheae and in the genus *Salvia* (Takano and Okada, 2010; Walker and Sytsma, 2007; Walker et al., 2004). In contrast, low sequence divergence has been found in both chloroplast and ribosomal regions in *Echinacea*. For instance, the sequences of *trn*S and *trn*G loci displayed low diversity within *Echinacea* and among outgroup species (Flagel et al., 2008). Moreover, sequence divergence of ITS1, ITS2 and 5.8S regions was low among a number of *Echinacea* species and several species were found to have identical ITS-2 sequences (Urbatsch et al., 2000). This poor phylogenetic resolution may be due to a combination of effects such as incomplete lineage sorting and hybridization between existing populations (Flagel et al., 2008). Previous studies have also found that the high incidence of hybridization, introgression and (allo) polyploidity in land plants has made more difficult the development of a barcode for plants than in animals (Chase et al., 2005; Fazekas et al., 2009). In order to improve the barcoding success in plants additional barcoding sequences not only from the plastid DNA but also from nuclear genes may provide improved species resolution. For instance, nuclear genes can provide a more reliable assessment of hybridization than plastid DNA that is uniparentally inherited (Alvarez et al., 2008; Fazekas et al., 2009). In addition, it is very unlikely that any chloroplast loci could be linked to a gene responsible for an important agronomical trait, since these genes are usually found in nuclear genome.

DNA-based analyses in particular PCR and sequencing-based methods are being extensively used for authentication of medicinal herbs and now are even being commercialized and patented. For instance, a SCAR marker that was used to distinguish medicinal plants such as *Panax ginseng* from *Panax quinquefolius* and adulterants was patented in the US (Shaw et al., 2009). However, is important to emphasize that in order

to implement these DNA-based analyses at the regulatory context (legislation and pharmacopeias) for medicinal herb authentication, there are also general requirements that are needed. A first crucial step is to standardize a protocol that allows the extraction of high quality DNA (Li et al., 2011; Shaw et al., 2009). This first step could be problematic depending on the source of the material and how it is been processed (heat, boiled, sun-dried) which may difficult the extraction procedure. Secondly, it would be required to validate that the technique is capable of identifying the correct species. For instance, DNA barcoding was found to be efficient in the identification of species of the Polygonaceae family that are refereed as medicinal plants in the Chinese pharmacopoeia (Song et al., 2009a). Additionally, in the case of DNA barcodes it is necessary to have a representative amount of good quality sequences that can serve as references, since the accuracy of the identification depends on them. Furthermore, it is important that these reference sequences are compiled for different species in a reference database (Li et al., 2011). Finally, the method should be tested for amplification of contaminants if regions, such as ITS and rDNA, are used (Li et al., 2011). Currently, new quality control techniques have been included in Chinese Pharmacopoeia (2010 edition) which include PCR-based identification of animal derived medicines such as *Bungarus multicinctus* and it adulterants. The application of this method for identification improved accuracy and reduced the analysis time (Gao et al., 2011). Therefore, more DNA-based analyses are expected to be included in the identification of medicinal material at the regulatory level.

1.4.3.3 Hybridization-based methods

Nucleic acid hybridization is the process of joining two complementary strands of DNA. Among the hybridization methods the array-based methods are becoming popular for gene expression and also for authentication of medicinal herbs with the aim of accelerating analyses and reduce costs (Guo et al., 2009; Hudson and Altamirano, 2006; Sucher and Carles, 2008; Zhang et al., 2007). A microarray is a collection of hundreds of microscopic DNA spots (called probes or features) printed on a solid support. Each probe/feature contains a different immobilized DNA sequences suitable for hybridization with a complementary DNA that has been previously labeled (target) (Chavan et al., 2006). Fingerprints are then generated by scoring presence or absence of each hybridized feature. A unique set of present hybridized features on the array will give a fingerprint for the analyzed sample (Jayasinghe et al., 2009). DNA microarrays can overcome many of the limitations of the PCR-based fingerprinting methods. For instance, they provide fixed data features, removing the positional variation inherent in gel fingerprints. In addition, this technique is based on hybridization, ensuring that common elements are identical and genetically informative (Kingsley et al., 2002).

DNA microarrays can be employed on the authentication of herbal material. For example, comparative sequence alignment of the 5S-rRNA gene of 20 toxic traditional Chinese medicinal species revealed species-specific nucleotide sequences that were used to design and synthesize oligonucleotide probes that were immobilized in a silicon chip. The target sequences were amplified and labeled with fluorescent dye using genomic DNA as templates. After hybridization, stringent washing and scanning, quantification of the fluorescent images indicated unequivocal identification of the toxic

species analyzed (Carles et al., 2005). Similar studies have used species-specific specific oligonucleotide probes designed from the 18S rRNA gene sequences of *Panax* taxa (Zhu, 2008) and from 26S rDNA genes of several *Fritillaria* species (Tsoi et al., 2003) to develop a microarray useful for the identification of these medicinal species and derived drugs.

DNA microarrays have not been commonly used for *Salvia* and *Echinacea* authentication. Most of the microarray studies in these genera are focused on gene expression that aim to evaluate the effect of *Echinacea* and *Salvia* formulas and bioactive ingredients in animal tissue and in cell cultures (Hudson and Altamirano, 2006; Lee et al., 2008; Sertel et al., 2011; Wang et al., 2006; Yin et al., 2010). One of the few studies performed for authentication used a comprehensive detection method which combined multiplexed ligation-dependent probe amplification (MPLA) with microarray to identify *Echinacea angustifolia* from other medicinal and toxic plants used in the study (Barthelson et al., 2006). This method used a ligation reaction that joins two oligonucleotide probes that matched one continuous sequence coupled with PCR amplification. In this study the cytochrome P450 gene sequences were used to design probes for the MPLA assay.

Although DNA microarrays have been proven to be a useful and rapid tool for authentication, the main disadvantage is that in order to synthesize the oligonucleotide probes, specific sequences unique to the species or varieties analyzed have to be found (Heubl, 2010). This is a great disadvantage if medicinal herbs with few gene sequences in the public databases need to be authenticated.

1.4.4 Alternative microarray techniques

There are alternative array techniques that do not require previous DNA sequence information, which are explained on the following sections. **Table 1.4** compares the main characteristics of the previously discussed oligonucleotide array with these alternative array techniques.

1.4.4.1 Diversity Array Technology (DArT)

This technology called Diversity Array Technology (DArT), uses microarrays to detect DNA polymorphism at several hundred genomic loci without any previous DNA sequence information (Jaccoud et al., 2001). The DArT technology involves the development of a "Discovery array" from a pool of genomes representing the germplasm of interest (Gupta et al., 2008). This genomic DNA pool is subjected to a genome complexity reduction by restriction enzyme digestion followed by ligation of restriction fragments to adapters and subsequent amplification. After that, individual fragments are cloned, amplified and spotted on glass slides to construct the "discovery array" **(Figure 1.4 A)**. Labeled genomic representations of individual genomes that have undergone the complexity reduction are hybridized onto the discovery array **(Figure 1.4 B)**. After scanning, image and data analysis the polymorphic clones (called DArT markers) are assembled into a new "genotyping array". This new array is used to generate a whole-genome fingerprint of any organism or group of organisms belonging to the pool from which the array was constructed by scoring the presence or absence of hybridization to individual array elements (Gupta et al., 2008; Jaccoud et al., 2001).

Tabla 1.4. Comparison between the array-based techniques.

Parameter	Oligonucleotide microarray	DarT	Diversity SSH array	SDA
Array preparation	Species-specific sequences are designed and synthesized.	A genome complexity reduction is performed by restriction enzyme digestion followed by ligation and selective amplification.	After multiple pair-wise subtractions, the subtracted DNA fragments are isolated by cloning and amplified.	After a single subtraction the subtracted DNA fragments isolated by cloning and amplified
Target preparation	PCR amplified products	A genome complexity reduction is performed by restriction enzyme digestion followed by ligation and selective amplification.	Restricted DNA fragments	Restricted DNA fragments
Sequence information required	Yes	No	No	No
Percentage of polymorphism	NS	3 - 27%	40.6 - 46.8%	10.5 - 68%
Part of genome surveyed	Part of genome	Whole genome	Subtracted part of the genome	Subtracted part of the genome
Major application on study of herbal medicines	Used for: -Gene expression studies that aim to evaluate the effect of herbal medicinal formulas and bioactive ingredients. -Authentication of herbal material.	Fingerprinting of closely related *Eucalyptus* trees	Species identification	Used for: -Genotyping. -Inferring genetic relationships -Authenticate herbal material, including dried commercial samples.

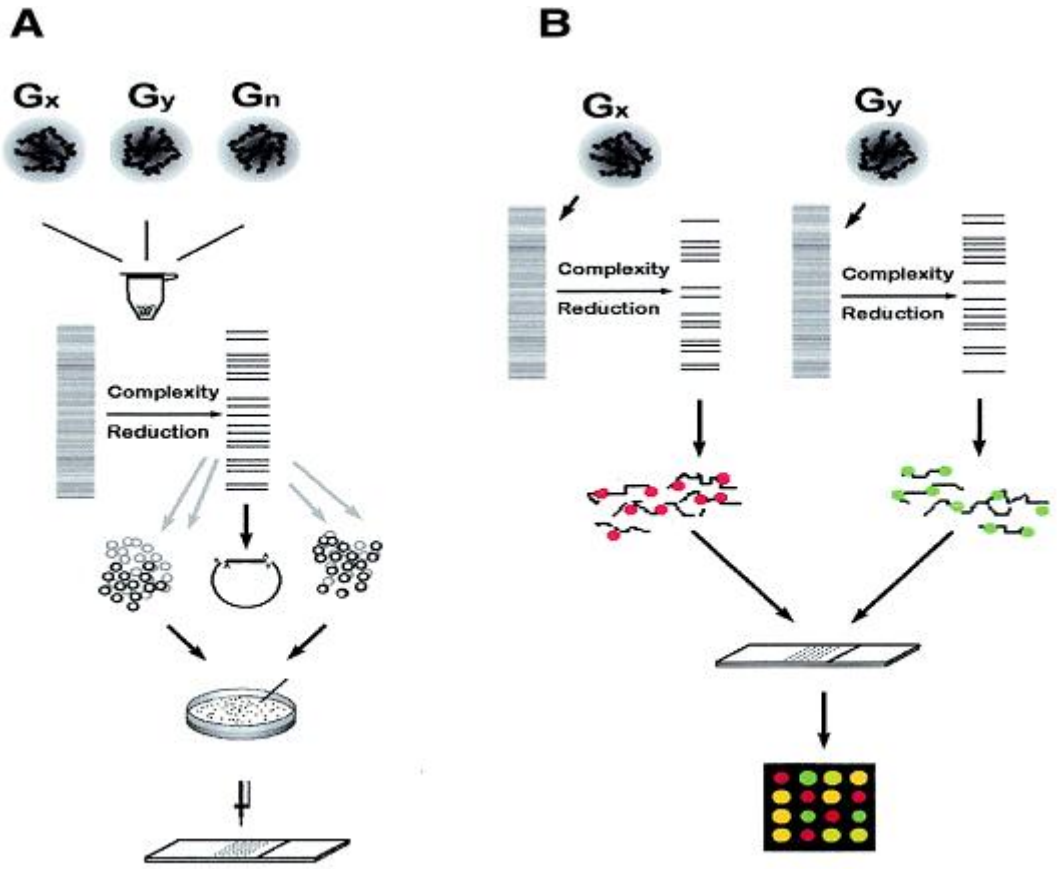


Figure 1.4. Process of DArT. (A) Generation of diversity panels. (B) Target preparation and hybridization of two contrasting samples (Jaccoud et al., 2001).

Diversity arrays are capable of detecting single base pair changes (SNPs) and insertion/deletion/rearrangements polymorphisms. The molecular basis of these polymorphisms was validated in a DArT study performed in *Arabidopsis thaliana*. In this study, the comparison of 107 sequences of ecotype Landsberg *erecta* (Ler) with the ecotype Columia (Col) genome sequence was able to clearly show that diversity arrays can detect SNP and insertion/deletion DNA polymorphisms in restriction sites (Wittenberg et al., 2005). In addition, if a methylation sensitive enzyme such as *Pst*I is used it may additionally identify polymorphisms due to DNA methylation (Gupta et al., 2008; Wittenberg et al., 2005).

This technique has been successfully used for genotyping of several crops such as Pigeonpea (Yang et al., 2006b), cassava (Xia et al., 2005), and rice (Xie, 2006) and also has been used to create genetic linkage maps for crops such as barley (Wenzl et al., 2004), wheat (Akbari et al., 2006) and sorghum bicolor (Mace et al., 2008). DArT also has been proven useful to study non-model plants. For instance it has been used to fingerprint *Eucalyptus grandis* (Lezar et al., 2004) and to perform evolutionary studies on the diploid fern *Asplenium viride* and the haploid moss *Garovaglia elegans* (James et al., 2008).

The main disadvantage of DArT is the method of genome complexity reduction used. This method is not very effective, since a large number of homologous/monomorphic sequences remain present in the array, which will affect the level of polymorphism (Li et al., 2006). The polymorphic rates reported for barley (2.9-10.4%) (Wenzl et al., 2006), cassava (9-14%) (Xia et al., 2005), wheat (5.3-9.4%) (Akbari et al., 2006) are relatively low which implies that a higher number of DArT clones need to be screened in order to find polymorphic markers. In addition, a high repetitive DNA content may result in cross-hybridization (binding of unspecific fragments in the probe), which can mask potential polymorphisms (Lezar et al., 2004).

1.4.4.2 Subtraction Suppression Hybridization (SSH)

An alternative to DArT is to use a technique called Subtraction Suppression Hybridization (SSH). Although SSH was developed for gene expression studies (Diatchenko et al., 1996), it has also proved to be useful for screening DNA polymorphisms and species-specific sequences from whole gDNA. Firstly, this

technique will be explained on the basis of its ability to identify differentially expressed genes (Diatchenko et al., 1996) and then their uses to screen for specific sequences from whole gDNA will be highlighted (Li et al., 2004).

SSH is used to selectively amplify target cDNA fragments (genes differentially expressed in two samples) and simultaneously suppress non target DNA amplification (genes expressed in both samples). This selective amplification and suppression is performed irrespective of the level of expression and in the absence of sequence information (Gadgil et al., 2002). The differentially expressed cDNAs that want to be identified are present in the "tester" cDNA but absent (or present at lower levels) in "driver" cDNA (Diatchenko et al., 1996). The process of subtraction has five important steps which are represented also in **Figure 1.5**.

- **Digestion of pooled cDNA**. Each cDNA is digested with a four-base cutting restriction enzyme (usually *Rsa*I since it generates large average size fragments of aprox. 600bp) to generate shorter cDNA fragments with a blunt end.

- **Adaptor ligation.** The cDNA of tester is divided in two samples, each ligated to a different adaptor at the 5' end of each fragment.

- **First hybridization.** An excess of driver is added to each of the adaptor-ligated tester samples. The sample is heat-denatured and allowed to anneal. In this step, the reanneling process generates homo-hybrid cDNA (b) and hetero-hybrids (c) cDNAs. The hetero-hybrids are formed from "common" non target cDNA

(present in tester and driver). Due to the faster reannealing of homo-hybrid cDNAs, the ss cDNA tester fraction (a) is normalized (concentrations of high and low abundance cDNA become roughly equal).

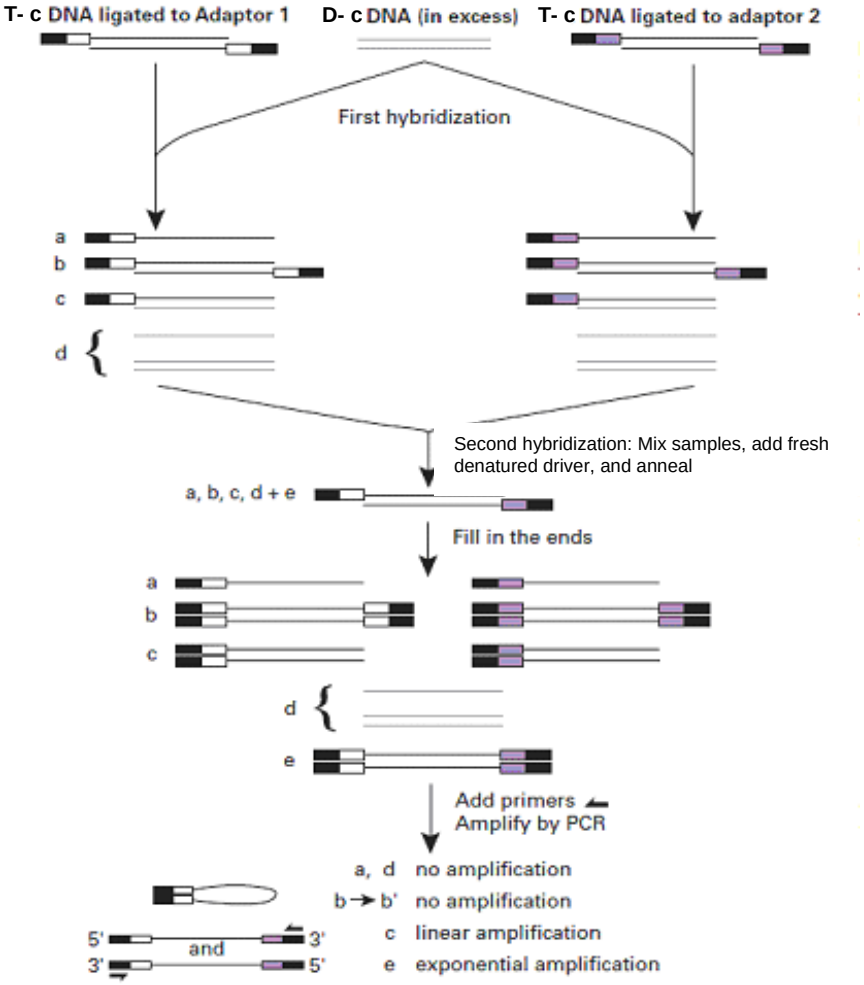


Figure 1.5. Abbreviated description of the SSH method. Solid lines represent the digested tester (T) or diver (D). Figure adapted from Clontech PCR- Select ™ cDNA Subtraction Kit User Manual (Protocol #PT1117-1, www.clontech.com).

- **Second hybridization.** The two end products of first hybridization are mixed together and additional excess of single stranded driver is added. In this step only the normalized ss tester cDNA are able to reassociate and form (b), (c), and (e) hybrid which has different adapter sequence at the 5'-ends.

- **PCR amplification.** After hybridization reactions, an extension reaction is performed to fill in the sticky ends of the molecules for primer binding. Only the molecules in which the two strands have different adaptors (e) can be exponentially amplified by PCR using a pair of primers which bind to the outer part of the adapter 1 and 2 respectively. Type b molecules, which contain complementary sequences on the ends, should be suppressed during the primer annealing step where the hybridization kinetics favors the formation of "panhandle-like" structures, which prevents the primer annealing and further extension (suppression effect). Finally, type (a) and (d) molecules do not contain primer binding sites and type (c) molecules are only linearly amplified.

As mentioned above SSH has also proved to be useful for screening DNA polymorphisms and species-specific sequences from gDNA (Li et al., 2006; Li et al., 2004). For instance, Li et al. (2006) used pair-wise DNA subtractions between four *Dendrobium* species in order to obtain differential gDNA fragments that were cloned, amplified and spotted as probes on positively charged nylon membranes to generate the diversity SSH arrays. After hybridization of DIG- labeled genomic DNA of the different *Dendrobium* species and image analysis, it was possible to identify the polymorphic and

species-specific probes. The percentages of polymorphic probes between two species varied from 40.6% to 46.8% **(Table 1.4)**. However, the main disadvantage of this technique is the use of multiple subtractions between the species analyzed; a total of four subtractions had to be performed in order to fingerprint only six *Dendrobium* species. Therefore, it is clear that this method is costly and time consuming if numerous species within a large genus such as *Salvia* were to be analyzed.

1.4.4.3 Subtracted Diversity Array (SDA)

A novel technique called 'Subtracted Diversity Array' (SDA) developed by Jayasinghe et al. (2007) combines an alternative SSH with a high-density microarray, increasing the chances to find polymorphic features.

The first prototype SDA was constructed from a pooled genomic DNA library of 49 angiosperm species, from which pooled non-angiosperm genomic DNA was subtracted. The suppression subtraction hybridization (SSH) involved the development of gDNA representations from angiosperms (tester) and non-angiosperms (driver), double-digestion of these representations with the restriction enzymes *Alu*I and *Hae*III, ligation of adaptors, two hybridization rounds and suppression PCR in order to amplify the sequences specific to flowering plants. These amplified products were then cloned producing a DNA library of 376 clones. Subsequently, the library inserts were PCR-amplified and precipitated in order to construct the array enriched with potential polymorphic angiosperm specific sequences (Jayasinghe et al., 2007). The main advantage of the SDA over the diversity SSH array of Li et al. (2004) is that instead of

making pair-wise subtractions between the species; genomic DNA representations are pooled and a single subtraction is performed using the pooled driver and tester.

In order to prove the usefulness of the SDA in genotyping, 28 angiosperm species were divided into six clades (Eumagnoliids, Caryophyllids, Rosids, Monocots, Ranunculids and Asterids). For each clade, equivalent DNA from each species was pooled, and then each pooled DNA was double-digested with *Alu*I and *Hae*III, labeled and hybridized to the SDA. Different patterns of positive features were found for each clade, including numerous features unique to each clade (Jayasinghe et al., 2007). Additionally, a second study found that this prototype SDA was able to successfully genotype to the family level and to the species level with some minor exceptions. Moreover, this second study also showed that SDA is capable to accurately genotype species not included in its construction (Jayasinghe et al., 2009). Therefore, this novel technique is capable of genotyping single or mixed genome representations from a wide range of plants. This is a significant advantage over diversity SSH arrays which are used to genotype only a narrow group of species (Jayasinghe et al., 2007).

The SDA hybridization results were also used to perform hierarchical clustering analysis by converting the signal to background ratio into binary data. The dendrograms generated by the two previous SDA studies were found to bear a strong resemblance to the phylogenetic tree established by the Angiosperm Phylogeny Group (APG) (Bremer et al., 2009). For instance, the angiosperms specific SDA was useful in classifying different families with the respective clade and species within their correspondent

families (Jayasinghe et al., 2009). This usefulness of the SDA to infer genetic relationships could be an important application of SDA in addition to the genotyping.

Furthermore, this-angiosperm specific SDA was found sensitive enough to discriminate between two ginseng species, *Panax ginseng* and *P. quinquefolius*, from DNA extracted from dried root tissues. In addition, it was able to detect deliberate adulteration of 10% *P. quinquefolius* in *P. ginseng* herbal preparations (Niu et al., 2011b). Therefore, this SDA could be used for authentication purposes of commercial samples.

Another important advantage of the SDA is the high level of polymorphism achieved (10.5-68%), which is higher than the typically achieved for DArT (3-27%), as can be seen in **Table 1.4**. One of the possible reasons for the increase in the level of polymorphism may be the use of frequent cutting restriction enzymes like *Alu*I and *Hae* III that recognize 4bp sequences, generating an increased number of shorter fragments compared to the 6bp cutters used in DArT arrays (such as *Pst*I and *Eco*RI). The presence of a higher number of shorter sequences produces higher numbers of selective nucleotides that can be used for selective amplification, which gives a better representation of the genome and an increased probability of finding polymorphisms (Niu et al., 2011a). Another possible reason for the increased polymorphism may be the subtraction process which by enriching the array for different or unique sequences may have increased the chances to find polymorphic features. For instance, in the study performed by James et al. (2008), standard DArT complexity reduction was compared with SSH, in order to construct a library for *Asplenium viride* and *Garovaglia elegans*,

after the genotyping of individual samples it was found that the percentage of polymorphism was higher for markers derived from SSH. For instance, for *Asplenium* the frequency of polymorphism was 8% for the diversity array constructed with the subtracted library compared to 4% from un-subtracted library. Therefore, the substantial enrichment for polymorphic sequences may be a result of elimination of highly conserved gDNA sequences through SSH.

The prototype SDA has also some disadvantages. For example, two closely related species, *Glycyrrhiza glabra* and *G. uralensis*, clustered further apart than expected in a dendrogram constructed in the second study performed by Jayasinghe et al. (2009). These results suggest that the discriminatory power of the prototype SDA was not enough to differentiate among closely related species or it could also suggest that the stringent data quality control employed in the analysis could have eliminated useful information from the data set (Jayasinghe et al., 2009). Therefore, new SDAs with higher discriminatory power should be developed if species and populations of the genus *Salvia* and *Echinacea* need to be analyzed; also a less stringent data analysis could be beneficial.

1.5 GENERAL CONCLUSION

Although *Salvia* and *Echinacea* have important medicinal and commercial value, correct identification of the commercial species is challenging. In both of these genera, misidentification among closely related species is common due to their morphological

similarity and the fact that hybrids between the closely related species occur spontaneously or from breeding programs. This morphological similarity has made it difficult to find characters that allow an accurate identification by macroscopic and microscopic methods. As a result, chromatographic methods are commonly used for quality control purposes on both of these genera; however, this technique is dependent on genetic, environmental, developmental and biological factors. DNA-based molecular marker methods are a potential technique for authentication of these two genera since they are not affected by environmental or developmental factors. However the genotyping of *Salvia* and *Echinacea* have raised some challenges since they are not model plants and there is a lack of genomic resources in public databases. PCR–based techniques such as RAPD and AFLP markers have the advantage that do not need previous knowledge of sequence, however they are based on gel electrophoresis, where correlating bands on gels for the allelic variants can lead to inaccurate interpretations (Jaccoud *et al.*, 2001). Sequencing-based methods have been used for species identification of these two genera, however low sequence divergence has been found in both chloroplast and ribosomal regions in *Echinacea* (Flagel et al., 2008).

SDA could be a potential technique for fingerprinting *Salvia* and *Echinacea* since it does not require previous DNA sequence information and has shown to be capable of isolating highly polymorphic sequences unique for the taxa under study. In addition, SDA poses significant advantages over similar techniques. For example, SDA needs only a single subtraction rather than pair-wise subtractions used for the diversity SSH array, also higher level of polymorphism (10.5-68%) has been achieved in comparison to the reported for DArT (3-27%). However the discriminatory power of the angiosperm

specific SDA constructed by Jayasinghe et al. (2007) was not enough to differentiate among closely related species. New SDAs specific for *Salvia* and *Echinacea* may have higher discriminatory power, which could improve the ability of SDA to genotype closely related species and even different accessions or populations within species for these two genera.

1.6 RATIONALE OF THE THESIS STUDY

DNA microarrays have been employed for the authentication of medicinal herbs such as *Panax, Fritillaria, Aconitum, Pinellia and Dendrobium*. However, this technique has not been widely applied for authentication of several other medicinal herbs such as *Salvia* and *Echinacea.* The main disadvantage of previous constructed microarrays is that in order to synthesize the oligonucleotide probes, specific sequences unique to the species or varieties analyzed have to be found. A new technique called 'Subtracted Diversity Array' has been used for efficient fingerprinting of angiosperms without prior knowledge of sequence information. This SDA was capable to fingerprint medicinal plants to the clade and family level. The aim of this project is to construct new SDAs to use them for fingerprinting species and populations of *Salvia* and *Echinacea.* In addition, it aims to validate if this technique has the potential to identify possible molecular markers that could be species-specific or that could be associated to important agricultural traits such as production of bioactive compounds.

The specific objectives of this study were:

1. Construct two SDA enriched for polymorphic DNA sequences, one for *Salvia* and other for *Echinacea*.
2. Fingerprinting of *Salvia* and *Echinacea* species.
3. Fingerprinting of geographical populations of *S. miltiorrhiza*.
4. Characterize (sequence) the DNA polymorphic fragments obtained from the arrays, since they could be potential molecular markers.
5. Correlate data from agronomic traits of previous studies with the microarray data in order to identify possible molecular markers associated to agronomic traits.

CHAPTER 2

Fingerprinting of *Salvia* species using the *Salvia* Subtracted Diversity Array

2.1 INTRODUCTION

This chapter describes how suppression subtractive hybridization between a pool of ten *Salvia* species and a pool of non-angiosperm and angiosperms (excluding the Lamiaceae) was used to enrich selectively the Subtracted Diversity Array (SDA) with polymorphic and divergent DNA sequences of the genus *Salvia*. Additionally, it is described how this *Salvia*-specific array was the used for fingerprinting several members of this genus.

Salvia (Lamiaceae) is an important genus with approximately 1000 species distributed widely in many regions of the world including the Mediterranean area, southern Africa, Central and South America, and Asia (Walker et al., 2004). As described in the literature review, several species of *Salvia* are known for their medicinal and culinary purposes (Topcu, 2006). However, several of the commercially important *Salvia* could be misidentified or adulterated with closely related species. For example, most of the commercial dried sage imported into North America not only consists of *S. officinalis* L. but it is often mixed with *S. fruticosa* Mill. since *S. officinalis* has a slow growth rate in winter months (Dudai et al., 1999). Another example is found in the commercialization of the root and rhizome of *S. miltiorrhiza* Bunge (commonly known as Danshen), where many of the morphologically similar species are used and traded under this same name

in different regions of China (Li, 2008). Therefore, accurate species identification is necessary since many of these related species could not have the same bioactive components and thus their pharmacological activity could differ from the correct one.

Commonly the correct species identification of the herbs has been performed by morphological or chemical techniques (Techen et al., 2004). However, techniques in molecular biology have given rise to new possibilities for identification based on the genetic composition of the plant which is unaffected by environmental, developmental and biological factors (Cai et al., 2002; Chan, 2003). For instance, the Random Amplified Polymorphic DNA (RAPD) markers have been extensively used to fingerprint different species of *Salvia* (Böszörményi et al., 2009; Bruna et al., 2006; Echeverrigaray and Agostini, 2006; Skoula et al., 1999). Also, chloroplast and mitochondrial DNA regions have been amplified and restriction digested (Polymerase chain reaction restriction fragment length polymorphism PCR-RFLP) to genotype *Salvia* species endemic to the Mediterranean region (Karaca et al., 2008). Furthermore, the sequences of the internal transcribed spacer (ITS) region have been amplified, sequenced and aligned in order to genotype different Chinese *Salvia* (Xu et al., 2009). However, these techniques have their disadvantages. For example, PCR–based techniques are based on gel electrophoresis, which is time consuming, and correlating bands on gels for the allelic variants is difficult and can lead to inaccurate interpretations (Jaccoud et al., 2001). In addition, chloroplast and ITS regions have not always been found polymorphic among closely related *Salvia*. For instance, high sequence similarities were found among these DNA regions of some Japanese *Salvia*, which made it difficult to determine evolutionary relationships among them (Takano

and Okada, 2010). Therefore, there is a need for species-specific genetic markers in *Salvia* that could be useful not only for the identification of closely related species but also that could be correlated to chemical profiles of the species studied in order to find molecular markers linked to desirable agricultural traits such as oil or bioactive compound content.

SDA (Subtracted Diversity Array) has shown to be able to isolate genomic DNA (gDNA) sequences specific for the taxa under investigation. Previously, an SDA was constructed using gDNA representations from angiosperms and non-angiosperms to isolate sequences specific to flowering plants (Jayasinghe et al., 2007). The array successfully genotyped to the clade and family-level and also was able to distinguish down to species level with minor exceptions (Jayasinghe et al., 2009). Furthermore, this SDA was used to discriminate among two ginseng species, *Panax ginseng* and *P. quinquefolius*, the latter being a common adulterant of *P. ginseng* herbal preparations (Niu et al., 2011b). Therefore, SDA could be a potential technique to fingerprint *Salvia* since it does not need previous DNA sequence information and has shown to be capable of differentiate between closely related species.

This chapter explores the potential of SDA to fingerprint *Salvia* species. The objectives of the experiments described in this chapter were: (1) to generate a SDA enriched for polymorphic and divergent DNA sequences for *Salvia* (2) to evaluate the potential of the SDA to fingerprint *Salvia* species, and (3) to identify species-specific sequences that may be potential molecular markers for species identification.

2.2 MATERIALS AND METHODS

2.2.1 Collection of plant material

In order to obtain a gDNA representation for the subtraction a total of 151 species including angiosperms and non-angiosperms were sourced **(Table 2.1)** for DNA extraction. Non-angiosperms were collected from Toolangi State Park, Victoria and identified (Duncan and Isaac, 1994). Angiosperms were obtained only from verified nursery species; a total of 126 species were sourced to represent all angiosperm clades.

In addition, a total of ten *Salvia* species (42 plants) were sourced to represent the different centres of diversity around the world. *Salvia miltiorrhiza* and *S. sinica* Migo plants were obtained from seeds which were used in a previous study (Li et al., 2009a). The other *Salvia* species were obtained from verified specimens from various plant nurseries **(Table 2.1)**.

2.2.2 Construction of the *Salvia* Subtracted Diversity Array

The construction of the SDA for *Salvia* is summarized in the **Figure 2.1**.

2.2.2.1 DNA extraction and development of tester and driver pools

Total DNA was extracted from fresh leaves using a modification of the standard CTAB procedure (Doyle and Doyle, 1987). Approximately 0.5 g of leaves was ground with liquid nitrogen to a fine powder. The powder was dissolved in 5 ml of CTAB Buffer (3% CTAB, 100 mM Tris-HCl pH 8.0, 20 mM EDTA pH 8.0, 1.4 M NaCl), 1 ml of 10% PVP and 1.2 ml of 10% CTAB.

Table 2.1. Description of the angiosperm and non angiosperm species used for DNA extraction and development of genome representations.

REPRESENTATIONS	SPECIES		
NON ANGIOSPERMS (25 species)	*Adiantum raddianum*	*Dicksonia antarctica*	*Riccardia eriocaula*
	Azolla sp.	*Equisetum hyemale*	*Selaginella* sp.
	Blechnum chambersii	*Ginkgo biloba*	*Sphagnum australe*
	Blechnum fluviatile	*Grammitis billardieri*	*Sticherus tener*
	Bryum billardieri	*Hymenophyton flabellatum*	*Thuidium* sp.
	Catagonium nitens	*Marchantia* sp.	*Weymouthia cochlearifolia*
	Cyathea cooperi	*Microsorum pustulatum*	*Wollemia nobilis*
	Cyathophorum sp.	*Polystichum proliferum*	
	Dawsonia superba	*Racopilum cuspidigerum v*ar.*convolutaceum*	
MAGNOLIIDS (6 species)	*Cinnamomum verum*	*Illicium anisatum*	*Nymphaea gigantea*
	Houttuynia cordata	*Magnolia denudate*	*Peumus boldus*
MONOCOTS (23 species)	*Acorus calamus*	*Dioscorea polystacha*	*Polygonatum multiflorum*
	Acorus gramineus	*Fritillaria thunbergii*	*Ruscus aculeatus*
	Aloe vera	*Iris domestica* (syn. *Belamcanda chinensis*)	*Serenoa repens*
	Bambusa beecheyana	*Iris versicolor*	*Trachycarpus fortunei*
	Bletilla striata	*Lilium longiflorum*	*Zea mays*
	Coix lacryma-jobi	*Lomandra longifolia*	*Zephyranthes sp.*
	Colocasia esculenta	*Ophiopogon japonicus*	*Zingiber officinale*
	Curcuma longa	*Pinellia cordata*	

EUDICOTS NOT PLACED IN EITHER THE ROSIDS OR ASTERIDS SUBCLADES (19 species)	*Aconitum carmichaelii* *Aquilegia* sp. *Berberis fortunei* (syn. *Mahonia fortunei*) *Berberis japonica* (syn. *Mahonia japonica*) *Buxus sempervirens*	*Chelidonium majus* *Clematis hexapetala* *Clematis montana* *Clematis serratifolia* *Clematis songarica* *Dianthus caryophyllus* *Dianthus superbus*	*Eschscholzia californica* *Grevillea robusta* *Gypsophila oldhamiana* *Hamamelis virginiana* *Phytolacca acinosa* *Ranunculus* sp. *Rumex crispus*
ROSIDS (33 species)	*Abutilon theophrasti* *Agrimonia eupatoria* *Agrimonia pilosa* *Albizia julibrissin* *Alchemilla xanthochlora* *Althaea officinalis* *Armoracia rusticana* *Astragalus membranaceus* *Baptisia tinctoria* *Allocasuarina* sp. *Catha edulis*	*Citrus aurantium* *Citrus reticulata* *Crataegus monogyna* *Dichroa febrifuga* *Filipendula ulmaria* *Firmiana simplex* *Glycyrrhiza glabra* *Glycyrrhiza uralensis* *Gynostemma pentaphyllum* *Humulus lupulus* *Hypericum perforatum*	*Isatis tinctoria* *Oenothera biennis* *Oenothera odorata* *Oxalis pes-caprae* *Passiflora edulis* *Pelargonium sp.* *Poncirus trifoliata* *Rosa rugosa* *Ruta graveolens* *Sophora flavescens* *Urtica dioica*
ASTERIDS (45 species) Excluding Lamiaceae	*Achillea millefolium* *Adenophora potaninii* *Angelica archangelica* *Angelica dahurica* *Aralia chinensis* *Artemisia abrotanum* *Artemisia absinthium* *Artemisia lactiflora* *Artemisia pontica* *Atropa belladonna*	*Codonopsis thalictrifolia* *Codonopsis pilosula* *Coffea arabica* *Cynara scolymus* *Digitalis purpurea* *Eupatorium perfoliatum* *Ilex paraguariensis* *Impatiens* sp. *Inula helenium* *Lycium barbarum*	*Sambucus nigra* *Scrophularia ningpoensis* *Scrophularia nodosa* *Solidago canadensis* *Symphytum officinale* *Syringa vulgaris* *Tanacetum cinerariifolium* *Tanacetum parthenium* *Taraxacum officinale* *Tetrapanax papyrifer*

ASTERIDS	*Bacopa monnieri* *Camellia sinensis* *Centella asiatica* *Chamaemelum nobile* *Clerodendrum trichotomum*	*Petroselinum crispum* *Physalis peruviana* *Plantago major* *Platycodon grandiflorus* 'Apoyama' *Platycodon grandiflorus*	*Trachelospermum jasminoides* *Tussilago farfara* *Valeriana officinalis* *Verbascum thapsus* *Withania somnifera*
SALVIA **(13 species)**	**SUBTRACTION POOL** **(total of 42 plants)** *S. officinalis* [e] (five plants) *S. lyrata* [b] *S. elegans* [a] *S. sclarea* [d e g] *S. mexicana* [a] *S. runcinata* [c] *S. lavandulifolia* [e]	*S. sinica* [f] (five plants from Zhejiang province) *S. przewalskii* [f] *S. miltiorrhiza* [f] (5 plants from 5 different populations) Shandong province Shanxi province Henan province Hebei province *S. miltiorrhiza f. alba* (Shandong)	**NOT INCLUDED IN THE SDA DEVELOPMENT** *S. lanceolata* [c] *S. microphylla* [a] *S. fruticosa* [d e g]

[a] Native to Central America. [b] Native to North America. [c] Native to South Africa. [d] Native to North Africa. [e] Native to the Mediterranean. [f] Native to China. [g] Native to Central and western Asia

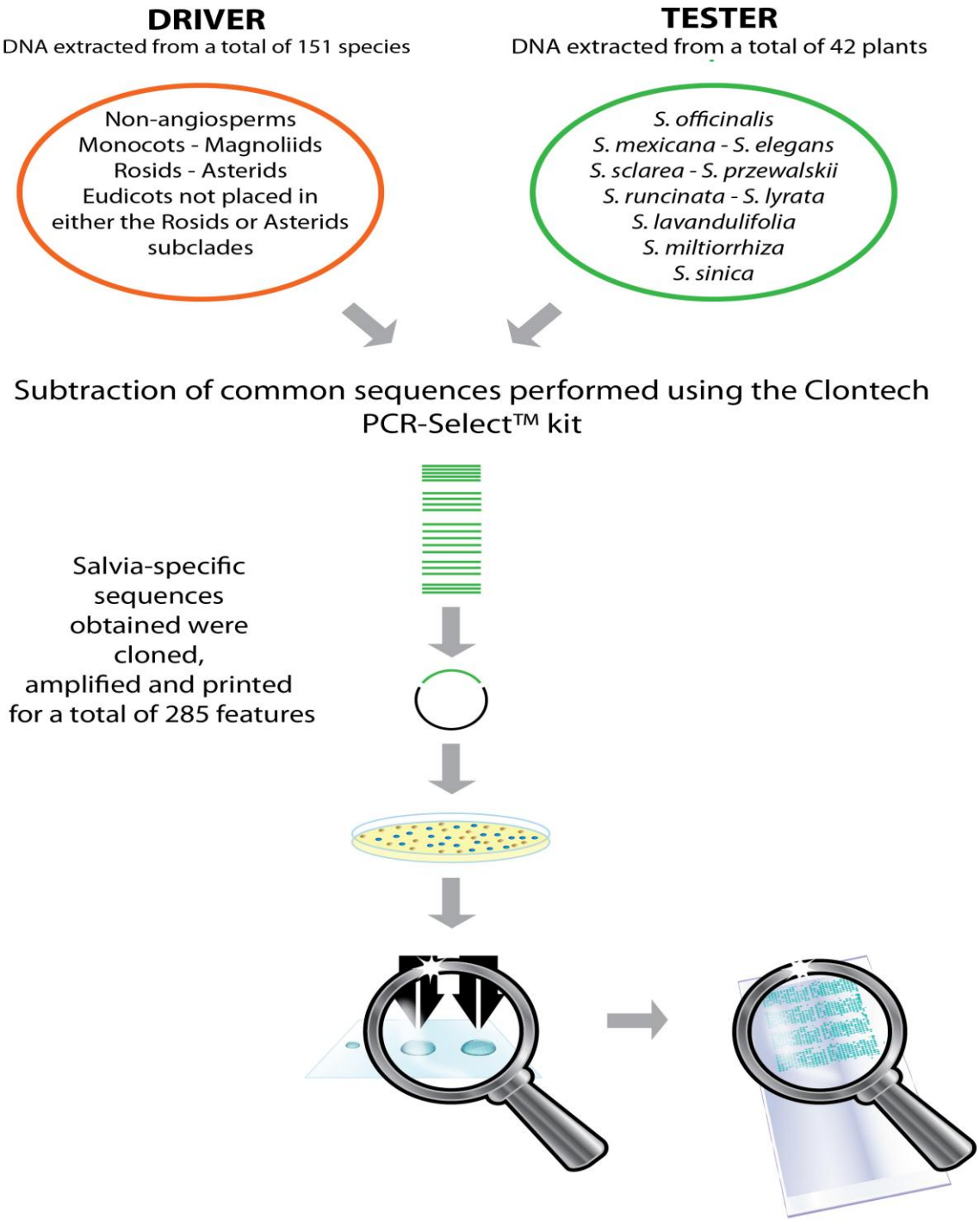


Figure 2.1. The process of construction of the *Salvia* SDA.

Subsequently, the mixture was incubated at 60^0C for 1 h and centrifuged at 13,000 rpm for 10 min. The supernatant obtained was further purified with a double chloroform extraction followed by precipitation with 7.5 M ammonium acetate and 100% ethanol. The precipitated DNA was resuspended in sterile water and subsequent clean up was performed by using the DNeasy® column of the DNeasy® Plant Mini Kit (Qiagen) following the protocol in the user manual.

All DNA samples were pooled based on the Angiosperm Phylogeny Group (2009) classification (Bremer et al., 2009) in order to obtain representations of the following seven groups: Salvia (Tester pool), non-angiosperms, Monocots, Magnoliids, Rosids, Asterids (excluding Lamiaceae), and Eudicots not placed in either Rosids or Asterids subclades (Eudicots and Core Eudicots) **(Table 2.1)**. About 10 µg of DNA was bulked for each representation, with each pool having equal amounts of genomic DNA per species. Subsequently, each pool was separately concentrated using the DNeasy® column of the DNeasy® Plant Mini Kit (Qiagen). The concentration and purity of the DNA pools were evaluated spectrophotometrically.

2.2.2.2 Suppression Subtractive Hybridization (SSH)

Subtraction was performed using the PCR-Select™ cDNA Subtraction Kit (Clontech), following the protocol in the user manual. The protocol was slightly modified to account for the double-stranded tester and driver as described below:

The *Salvia* pool (tester) was represented by equal amounts of DNA extracted from 42 different plants of which 25 were *Salvia miltiorrhiza* (5 plants from 5 different lines). The driver pool was obtained by bulking 700 ng of each representation with the exception of the *Salvia* pool **(Table 2.1)**.

In order to perform the subtraction, 4 µg of the driver and tester pool was digested overnight with *AluI* and *Hae*III (Fermentas) in a 60 µl reaction mixture containing 35 units of each enzyme. Digestion was verified by gel electrophoresis as described in the user manual. After that, digested DNA was purified by phenol:chloroform extraction, precipitated by ethanol-ammonium acetate and resuspended in 5.5 µl of sterile MilliQ water. Then, this digested *Salvia* pool was divided in two samples, which were individually ligated to a different adaptor (Adaptor 1 and 2R). Further, 0.3 ng of human skeletal muscle cDNA (control) was added to the tester before ligation. This was deliberately added in order to positively verify the ligation of adaptors in the *Salvia* pool.

In order to verify the ligation, a PCR was performed with primers specific for the adaptors (PCR primer 1) and the human skeletal muscle cDNA (G3PDH 3' and 5' primers). The ligation analysis was performed as follows:

Table 2.2 Setting up of the ligation analysis

	Volume in µl			
Component	***1***	***2***	***3***	***4***
Digested *Salvia* pool ligated to adaptor 1+ cDNA SM	1	1	-	-
Digested *Salvia* pool ligated to adaptor 2R+ cDNA SM	-	-	1	1
G3PDH 3' Primer	1	1	1	1
G3PDH 5' Primer	-	1	-	1
PCR Primer 1	1	-	1	-
Total volume (µl)	3	3	3	3

A PCR reaction performed with a successful ligation template would amplify a product of about 750 bp using G3PDH 3'Primer and PCR Primer 1 (Tubes 1 and 3). Products obtained in tubes 2 and 4 confirmed the presence of cDNA human skeletal muscle (SM). **Figure 2.2** shows a PCR product of about 750 bp for samples 1 and 3, therefore in both samples the ligation was successful. The human skeletal muscle cDNA was subsequently removed from the *Salvia* specific DNA during hybridizations by adding a total of 3.6 ng of human skeletal muscle cDNA in the driver.

Subsequently, two hybridizations were performed at 68^0C employing a tester: driver ratio of 1:30. During the first hybridization, the driver is added to each sample of the adaptor-ligated *Salvia* DNA. In this step the homologous sequences between the driver and tester were hybridized, leaving almost all of the *Salvia* specific DNA single

stranded. After that, the two sets of adaptor-ligated *Salvia* DNA from the first hybridization and a second portion of denatured driver are combined in one tube for the second hybridization. At the end of the second hybridization, *Salvia* specific DNA should be the only double stranded DNA with a different adaptor sequence at the 5'-ends.

Then, PCR was used to amplify exponentially *Salvia* specific DNA and nested PCR (secondary PCR) was used to reduce background and to amplify longer molecules of the enriched *Salvia* gDNA. Primary and secondary PCR were performed in a Perkin-Elmer GeneAmp PCR System 2400 with hot start as follows:

Primary PCR

- Incubate the reaction mix at 75^0C for 1 min to fill in the missing strands of the adaptors, creating the binding site for Primer 1.
- Add 0.5µl of Taq Polymerase on each tube.
- Incubate the reaction for 5 min at 75^0C
- Start thermal cycling

94^oC	2 min
32 cycles	
94^oC	30 sec
66^oC	45 sec
72^oC	1.5 min

72^oC for 5 min (final extension)

Secondary PCR

- Incubate the reaction mix at 94^0C for 3 min

- Incubate the reaction for 3 min at 80^0C and add 0.5ul of Taq Polymerase on each tube.
- Start thermal cycling

23 cycles:

94^{o}C	33 sec
68^{o}C	45 sec
72^{o}C	1.5 min

72^{o}C for 5 min (final extension)

Figure 2.3 shows the patterns obtained for the secondary PCR products of the subtracted *Salvia* pool and controls. It can be seen that the patterns are different for subtracted and unsubtracted samples which is the expected result in a successful subtraction.

2.2.2.3 Cloning of the subtracted sequences

Amplified products of the nested PCR were purified using the QIAquick PCR Purification Kit (Qiagen). Then, approximately 100 ng of the purified PCR products (*Salvia* specific DNA) was ligated into the pGEM®-T Easy vector (Promega) and transformed into heat-shock competent *Escherichia coli* JM109 (Promega) according to the user manual. Single colonies were grown overnight at 37^0C in LB medium supplemented with ampicillin, X-Gal and IPTG. White colonies were subcultured into LB/ampicillin broth and grown overnight. Positive transformation was determined by PCR amplification of the cloned insert using the nested primers from the subtraction kit. Plasmids containing cloned inserts which showed a single band were isolated from subcultured transformed cells using DirectPrep 96 Miniprep Kit (Qiagen). Finally, subcultured transformed cells were diluted in one volume of sterile glycerol and stored at -70^0C.

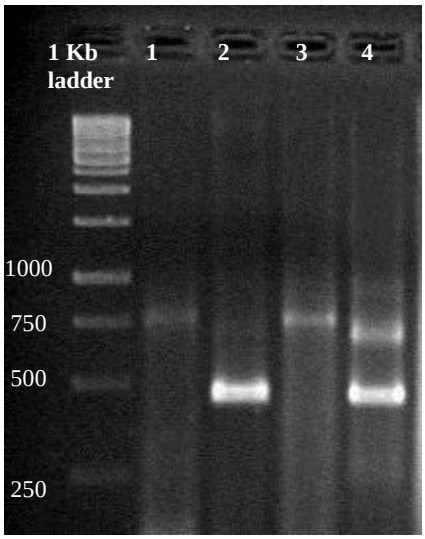


Figure 2.2. Results of the ligation efficiency analysis

Lane 1. PCR products using digested *Salvia* pool ligated to **adaptor 1** with added cDNA skeletal muscle as a template. **G3PDH 3' primer and PCR Primer 1**.
Lane 2. PCR product using digested *Salvia* pool ligated to **adaptor 1** with added cDNA skeletal muscle as a template. **G3PDH 3' and 5' Primers.**
Lane 3. PCR products using digested *Salvia* pool ligated to **adaptor 2R** with added cDNA skeletal muscle as a template. **G3PDH 3' primer and PCR Primer 1**.
Lane 4. PCR products using digested *Salvia* pool ligated to **adaptor 2R** with added cDNA skeletal muscle as a template. **G3PDH 3' and 5' Primers.**

1.5% agarose/EtBr gel.

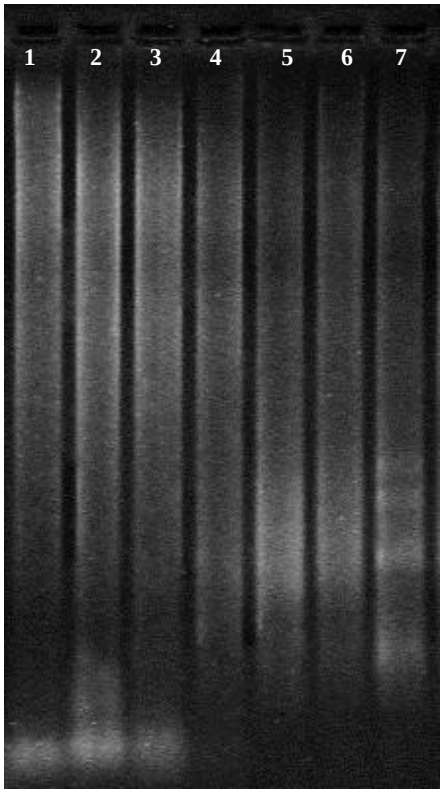


Figure 2.3. Secondary PCR products of the subtracted *Salvia* pool

Lane 1 and 3. Subtracted *Salvia* pool.
Lane 2. Subtracted skeletal muscle cDNA containing ΦX174/HaeIII-digested DNA
Lane 4 and 6. Unsubtracted *Salvia* pool.
Lane 5. Unsubtracted skeletal muscle cDNA containing ΦX174/HaeIII-digested DNA
Lane 7. PCR control subtracted cDNA provided with the kit.

1.5% agarose/EtBr gel.

2.2.2.4 Microarray construction and printing

The cloned inserts were PCR amplified from the corresponding plasmid using nested primers 1 and 2R (Clontech). The reaction mix contained 0.2 mM of each dNTP, 1.5 mM of MgCl2, 0.15 µM of each primer and 2 µl of purified plasmid. Amplification consisted of 30 cycles of 94^0C for 10 s, 68^0C for 30 s, extension at 72^0C for 1.5 min, with initial denaturation at 94^0C for 5 min and a final extension step of 72^0C for 5 min. After that, PCR products were precipitated in 96% ethanol and 3 M sodium acetate (pH 5.2). The precipitation was carried out at -20^0C overnight. The pellets obtained were washed with 70% ethanol, air dried and resuspended in 10µl of 50% DMSO. Finally, a total of 285 clones with inserts ranging from 250 bp to 1000 bp, were transferred into a 384-well plate (Genetix, Hampshire, UK) together with positive and negative controls **(listed on Appendix 1)**. Among the positive controls were three housekeeping genes (ribulose-1,5-bisphosphate carboxylase/oxygenase, ribosomal RNA, and chlorophyll a/b binding protein) sourced from *Cicer arietinum* (Coram and Pang, 2005).

The 285 clones together with the controls were printed on aminosilane-coated slides using a BioRobotics® MicroGrid II Compact arrayer (Genomic Solutions) at RMIT University, Australia. The parameters used to print are found in Appendix 2. Eight subarrays were gridded on a Corning® GAPS II coated slide (Corning Incorporated Life Sciences, Acton, MA). Each subarray was composed of 285 samples and 15 controls **(Appendix 1)**. A single printed slide was used to perform two hybridization experiments, where each hybridization reaction was tested with 4 subarrays. Following printing, the spotted DNA was rehydrated by steaming the printed slide surface and

dried over a heated block with the printed side up. The slides were then ultraviolet irradiated for 5 min and heated at 80^0C for 3 h.

2.2.3 Validation of the array

The SDA was validated by hybridizing separately the DNA from the *Salvia* and driver pools. The process of labeling and hybridization is summarized in the **Figure 2.4**.

2.2.3.1 Biotin labeling of target DNA

Salvia and driver pool (Target DNA) were digested with *AluI* and *Hae*III and the products were subsequently column purified (QIAquick PCR Purification Kit, Qiagen). Approximately 150 ng of purified digested DNA was labeled with Biotin-11-dUTP using the Biotin DecaLabel^{TM} DNA Labeling Kit (Fermentas, ON, Canada). The reaction was incubated at 37^0C during 20 h and no further purification was performed.

2.2.3.2 DNA Hybridization

Microarray slides were pre-hybridized for 40 min at 42^0C in a solution containing 5 x SSC, 0.1% SDS, 1% BSA and 25% formamide. Slides were rinsed with deionized water and dried using an air gun. Approximately, 30 ng of biotin-labeled sample was mixed with 17.5 µl of 2 x hybridization buffer (5 x SSC, 0.2% SDS, 50% formamide); 0.5 µl of 1 mg/ml Human Cot1 DNA (Invitrogen); 0.5 µl of 5 mg/ml PolyA (Sigma-Aldrich) and 0.5 µl of 10 mg/ml of salmon sperm DNA (Sigma-Aldrich). The mixture (made up to 35 µl with sterile water) was then denatured at 100^0C for 2 min and applied to the array under a 22x22-mm lifter slip (Grale Scientific, Australia).

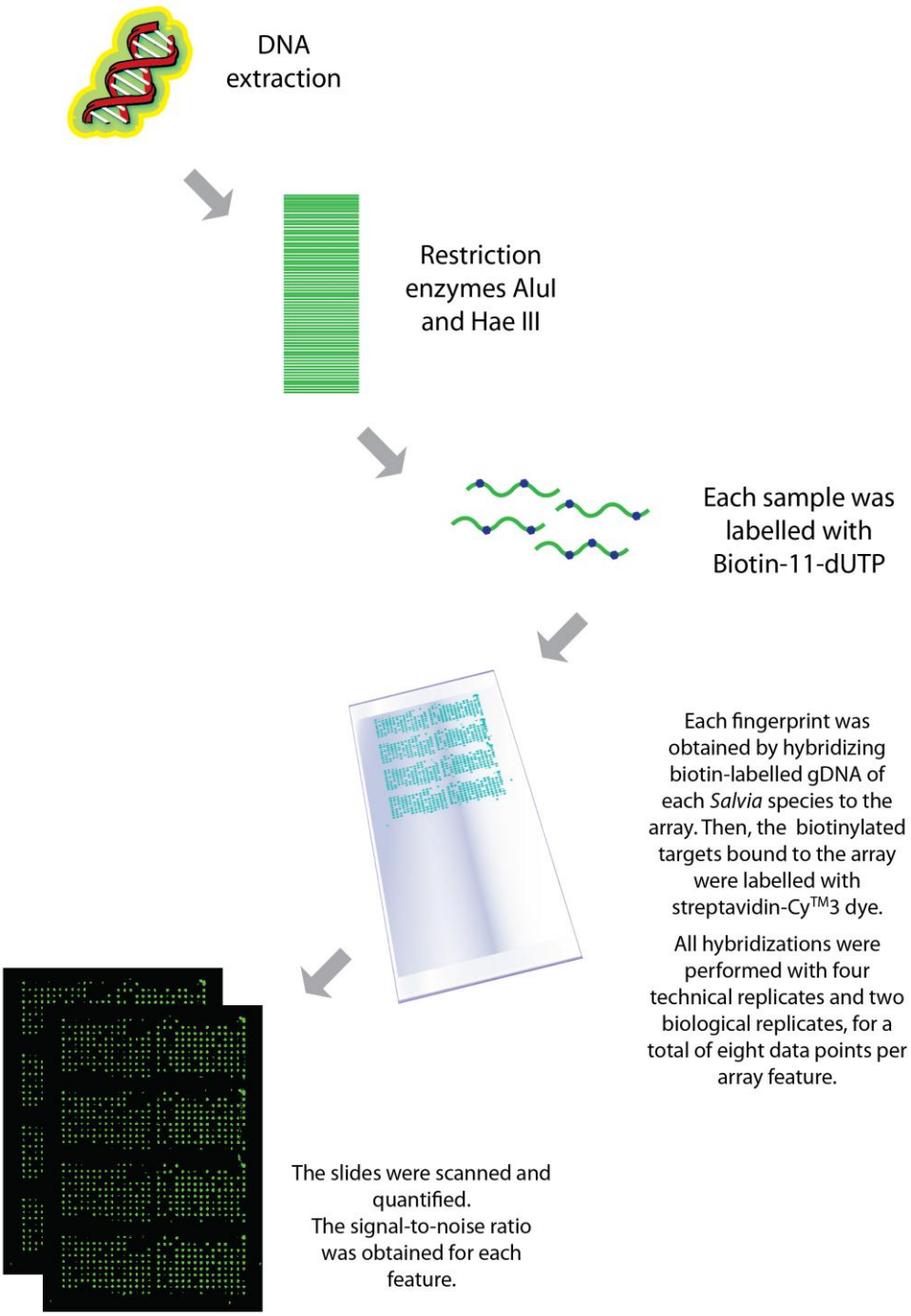


Figure 2.4. Process of biotin labeling and hybridization

Hybridization was performed overnight in a water bath at 42^0C in a waterproof and humidified hybridization chamber (Corning). All hybridizations were performed with four technical replicates (subarrays) and two biological replicates, for a total of eight data points per array feature. After hybridization, the coverslips were removed and the slides were washed twice in 1 x SSC, 0.1% SDS at 40^0C for 8 min, once in 0.1 x SSC, 0.1% SDS at 40^0C for 8 min and once in 0.1 x SSC at room temperature for 5 min. Then, detection of the biotinylated DNA targets bound on the array was performed by a protocol modified from Mirus Label IT® µArray® Biotin Labeling Kit (Mantri et al., 2011). Briefly, the slides were washed once in 6 x SSPE-T buffer at room temperature for 5 min. Buffer 6 x SSPE-T contains 300 ml of 20 x SSPE (175.3 g of NaCl, 27.6 g of $NaH_2PO_4 \cdot 2H_2O$ and 7.4 g of EDTA), 50 µl of Triton X-100 and 700 ml of MilliQ water. Immediately after the wash with 6 x SSPE-T buffer, the detection solution was applied to the wet surface of the slide and covered by a 25x60-mm lifter coverslip (Grale Scientific) to evenly distribute the solution. Detection solution was made of 200 µl of 6 x SSPE-T, 0.8 µl of 25 µg/ul of BSA and 0.5 µl of streptavidin-labeled Cy™3 dye (Amersham Pharmacia, Buckinghamshire, UK). Afterwards, the slide was incubated at 37^0C for 40 min in a waterproof hybridization chamber in the dark. Finally, the slides were washed three times in 6 x SSPE-T buffer for 5 min, rinsed with deionized water and dried with an air gun.

2.2.4 Fingerprinting of fifteen *Salvia* genotypes

Fifteen *Salvia* genotypes were fingerprinted, each fingerprint was obtained by hybridizing biotin labeled DNA of each *Salvia* species to the array. Labeling of the target and hybridization were performed as described in **Sections 2.2.3.1 and 2.2.3.2.**

However, the hybridization conditions were previously optimized in order to obtain a clear differentiation between species. For example, the hybridization temperatures were chosen after performing a gradient test of 5^{o}C (42 ^{o}C - 47 ^{o}C - 52 ^{o}C) with *S. officinalis*, *S. lavandulifolia*, *S miltiorrhiza* and *S. sinica* targets. Therefore, it was not until the hybridization and data analysis was performed with these three different temperatures that it was possible to determine that 42^{o}C was the optimal hybridization temperature for the *Salvia* array. Similarly, in order to optimize the amount of biotin-labeled sample different amounts of labeled DNA were used (1µg to 20ng). After hybridizing different amounts of the same target the reproducibility ranged from 0.85 to 0.95. This reproducibility was obtained after comparing the data sets using a linear correlation between them. This variability between the replicates was reduced (>0.91) hybridizing approximately the same amount of DNA (30ng).

2.2.5 Analysis of the *Salvia* array

2.2.5.1 Scanning and quantitation of spot intensities

Slides were scanned at 10 µm resolution and gain of 50 PMT using a Perkin Elmer array scanner. Images were captured and quantified with the ScanArray Express® Microarray Analysis System.

The quantitation was performed using the Easy Quantitation type. The printed spots on the image were matched to a designed grid **(Figure 2.5A)** which template specifications are as described in **Appendix 3**. Once the grid was fully adjusted to fit all the spots, the signal intensities were quantified using adaptive circle method and local background

A **B**

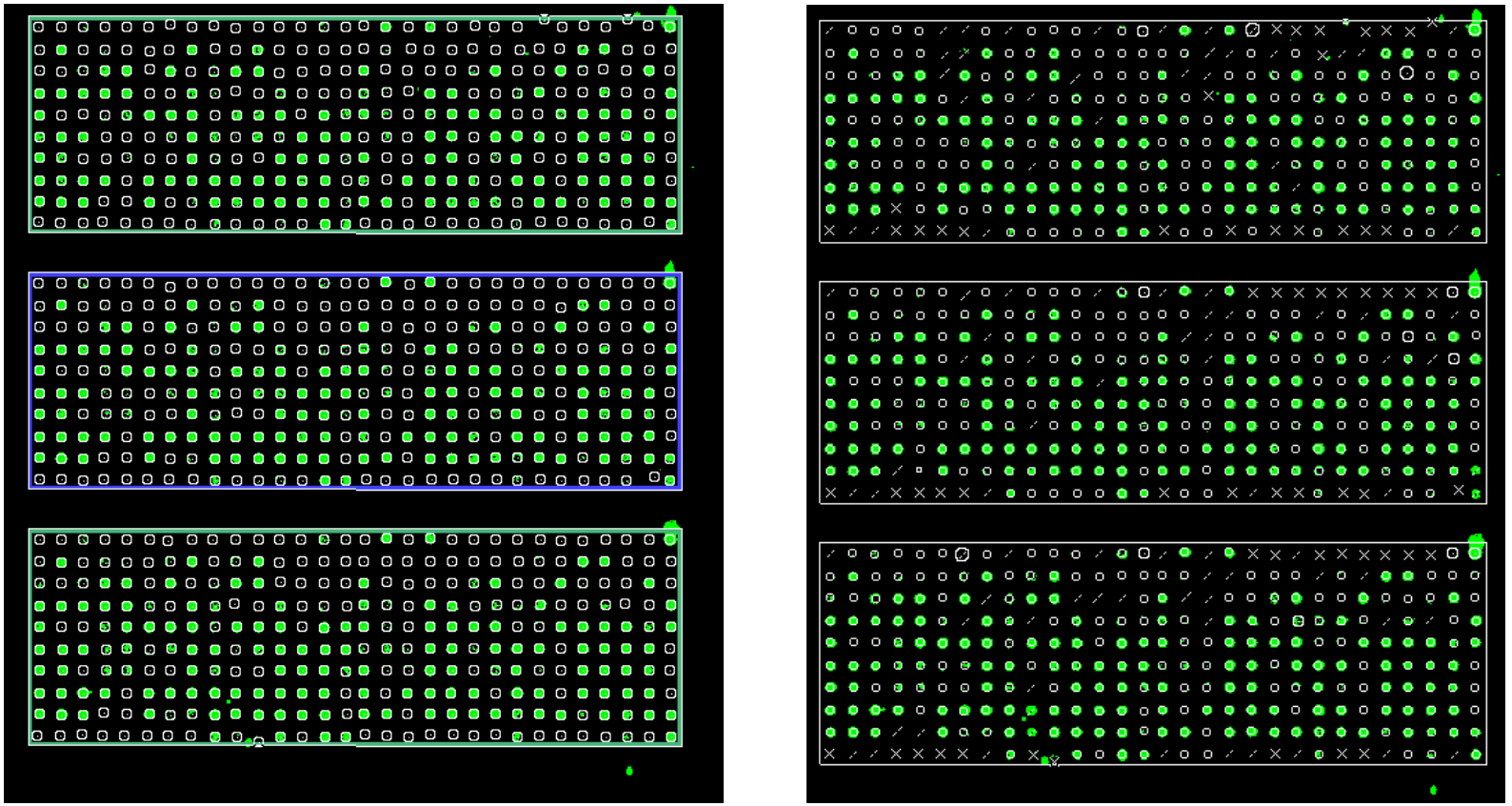


Figure 2.5. Quantitation of the scan images using ScanArray Express® Microarray Analysis System: A. Positioning of the grid onto the four subarrays. **B.** Flagged spots (X) that represent those with low quality/intensity.

was subtracted during quantitation. Abnormal spots that were not automatically flagged by the software were flagged manually; sometimes it was necessary to manually re-adjust the spot position by visual inspection **(Figure 2.5 B)**. A minimum signal-to-noise ratio of 7 was selected for the quality measurement. Finally, all the quantified data was exported to Microsoft Excel (Microsoft).

2.2.5.2 Data analysis

The following series of data transformations and subsequent statistical analyses are summarized in the **Figure 2.6**.

1. **Background correction:** The signal-to-noise ratio defined as:

$$\text{Signal-to-noise ratio} = \frac{\text{Mean Foreground - Mean Background}}{\text{Standard deviation of Background}}$$

was obtained for each spot by ScanArray Express®. This signal value was considered to have the most accurate background correction since it also accounted for variations in background intensity over the array.

2. **Omit flagged spots:** Automatically and manually flagged spots were omitted from analyses to ensure only high quality spots remained.

3. **Signal-to-noise ratio used for analysis:** Initially the scoring of the array was performed by converting the signal-to-noise ratio of each feature to binomial using the same analysis as described by Jayasinghe et al. (2007). The cut-off chosen was based on the hybridization signal of a specific control feature present in the array. This control was an aliquot of the enriched *Salvia*-specific sequences obtained from the subtraction process prior to cloning **(Appendix 1)**. In theory, if the subtraction was 100% efficient this control will act as a negative control for the driver target; however if a complete subtraction is not achieved, the signal-to-noise ratio obtained from that feature would represent the inherent background (unsubtracted sequences) given the experimental circumstances. Therefore, any signal-to-noise lower or equal to the signal-to-noise of this control was considered a negative spot. However, when this cut-off point was applied to the data obtained after fingerprinting the species, a significant number of features were not clearly differentiated as positive or negative; for example, for the four technical replicates two spots could be assessed as present and two as absent. Therefore, due to the lack of reproducibility of the binomial data, all the subsequent analyses were performed with the raw data of the signal-to-noise ratio. However, the efficiency of subtraction was calculated using binomial scoring in order to compare the efficiency of the subtraction with previous studies.

4. **Mean of signal intensity across the technical replicates:** A mean of the signal-to-noise ratio (signal intensity) between the four technical replicates was obtained for each of the 285 features.

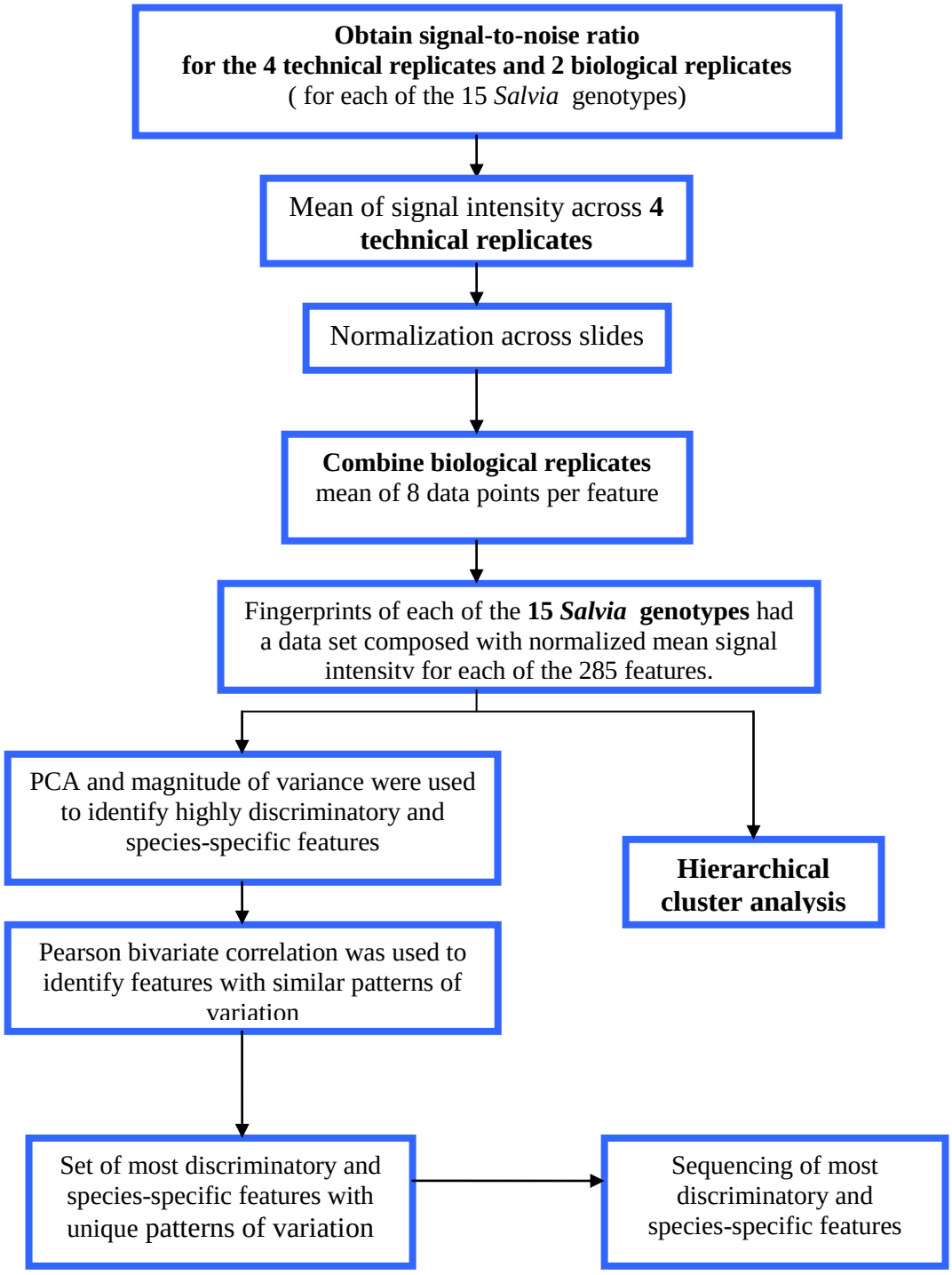


Figure 2.6. Flow-chart showing the data analyses performed to the signal-to-noise ratio obtained after scanning and quantitation of spot intensities.

5. **Normalization:** The data was then normalized across the slides using the following normalization ratio:

Normalization ratio = A / B, where:

A= Average signal intensity for all 285 features in all four technical replicates in **all hybridizations performed for all species.**

B = Average of the signal intensity of the four technical replicates across all 285 features **for a particular species.**

This ratio was used to normalize the signal of each feature across individual hybridization experiments.

6. **Combine biological replicates:** After normalization, the data of the biological replicates was combined to produce a single value per feature per species fingerprinted.

7. **Log transformation:** Logarithmic transformation was not necessary since the signal intensity was not significantly correlated with the variance.

2.2.5.3 Statistical analysis

1. **Histogram of signal intensities obtained after hybridization of tester and driver pools.** The raw signal-to-noise ratios (signal intensities) obtained for the 285 features after scanning and quantifying were background corrected, flagged and used to perform individual histograms for tester and driver targets. The histogram (MINITAB® Release 14.1) was performed in order to visualize the frequency of low and high signal intensities obtained for tester and driver targets.

2. **Principal component analysis of the full set of features:** The normalized mean data of the 285 features across the fifteen genotypes was input as the correlation matrix for PCA (MINITAB® Release 14.1). The variables of this matrix were the fifteen genotypes, thus the number of components to compute were 15. The output of this analysis displayed a score plot which was able to distinguish the features that accounted for most of the variability found across the genotypes.

3. **Variance of each feature across the genotypes:** The variances of the normalized mean signal across the fifteen genotypes fingerprinted were calculated for each feature in order to identify the features with the highest variances across the genotypes. This analysis was useful for identifying species-specific features which were not detected by PCA.

4. **Pearson bivariate correlation among the most discriminatory and species-specific features:** The normalized mean data of the most discriminatory and species-specific features were subjected to Pearson bivariate correlation (SPSS version 17.0). The variables for this analysis were the features chosen and a two tailed test of significance was performed. This correlation was performed in order to identify features with similar patterns of variation.

5. **Hierarchical cluster analysis of the fingerprints of 15 *Salvia* genotypes:** The normalized mean signal values of 285 features were transferred to PASW Statistics 18 to perform a hierarchical cluster analysis of the 15 *Salvia* genotypes. The dissimilarity dendrograms were generated using the average distance linkage between-groups and

Squared Euclidian metrics. The entire data set was used to perform the initial hierarchical cluster without excluding the features that hybridized with the driver target. These features were not excluded from the initial hierarchical cluster since its removal could eliminate some useful information from the data set, as it was found in previous studies where high stringent data analysis limited the establishment of relationships between closely species (Jayasinghe et al., 2009).

2.2.6 Sequencing of selected most discriminatory and species-specific features

The cloned inserts were re-amplified from the corresponding isolated plasmid using SP6/T7 primers. Amplification products were bidirectionally sequenced by Macrogen Inc. (Korea). Vector and primer sequences were removed and nucleic acid and protein homology searches were performed using blastN and blastX programs through the National Center of Biotechnology Information (www.ncbi.nlm.nih.gov).

2.3 RESULTS

2.3.1 Subtraction efficiency and validation of the microarray

Thirty-four (12%) positive features were found after hybridizing the driver target with the array. Theoretically, a complete subtraction should result in the absence of hybridization of the driver pool as all driver sequences are supposed to be eliminated; thus, these features that gave signal on both driver and tester represent the sequences that were not fully subtracted. Accordingly, the subtraction was able to isolate *Salvia*-

specific DNA sequences with 88% efficiency and the 12% of the features may represent non-subtracted sequences.

Furthermore, the histograms were performed in order to visualize the frequency of low and high signal intensities obtained for tester and driver targets and to find a cut-off point that could be used to differentiate signal intensities given only by *Salvia*-specific sequences (tester). These two histograms when superimposed **(Figure 2.7)** show that the signal intensities of the tester overlapped with that of the driver between 0 and 60. Therefore, features which signals were in this range could not be clearly differentiated as positive or negative, since there was not a clear cut-off point to separate the hybridization signal of *Salvia*-specific sequences from the hybridization signal of the species from the driver pool. Consequently, all subsequent analyses were performed with the raw data of the signal intensity.

2.3.2 Fingerprinting of fifteen *Salvia* genotypes and identification of the most discriminatory and species-specific features

Fingerprints for fifteen *Salvia* genotypes were obtained, which included fingerprints for thirteen species and two accessions of *Salvia officinalis*. Out of these thirteen, three fingerprints were of S. *lanceolata* Lam., *S. microphylla* Kunth and *S. fruticosa*, which were species not used in the construction of the SDA **(Table 2.1)**. Representative photographs of the fingerprints can be seen in **Appendix 4**.

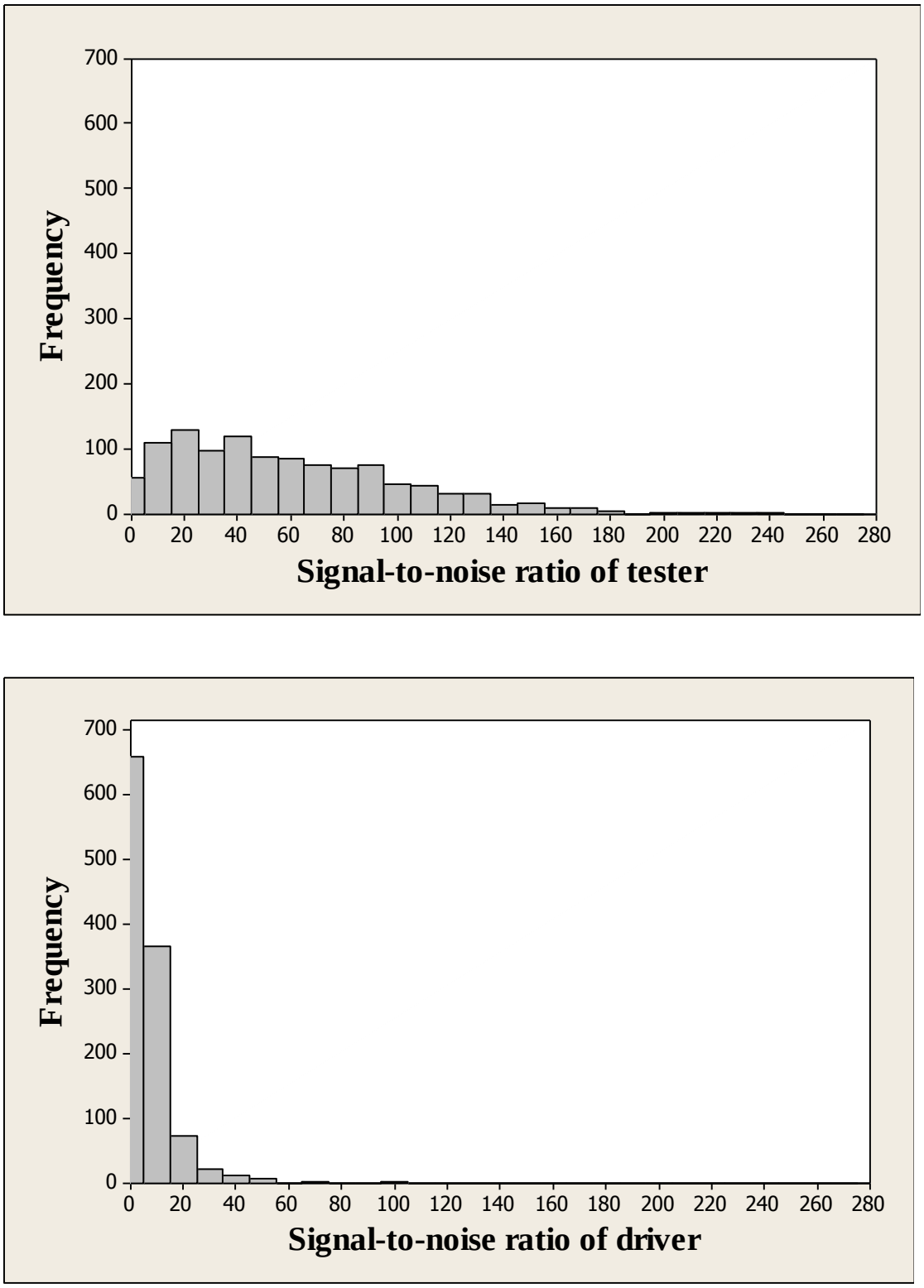


Figure 2.7. Histogram of the signal intensities obtained after hybridizations of the tester and driver pools.

A hierarchical cluster dendrogram constructed based on the 285 features grouped the fifteen genotypes into three distinctive clusters **(Figure 2.8)**. Cluster A contained species native to the Mediterranean region. Cluster B grouped all species native to Africa or the Americas together with *S. sclarea* L. and *S. fruticosa*, which are native to Europe, North Africa and Asia, and Cluster C grouped all *Salvia* species from China.

Further analysis with principal components (PCA) was performed in order to identify the most discriminatory DNA fragments capable of generating a fully resolved phylogeny of the genotypes analyzed. As it can be seen in **Figure 2.9**, a high percentage of variation (80.5%) may be explained by the first two components. The first principal component accounted for 71.1% of the variation and the second component explained only 9.4% of the variation. In addition, it was observed that most of the features clustered around zero with only a small number of features forming a loose cluster along the first component axis. Based on this analysis, only the four most distant features from zero on the X axis were chosen since the first component explains most of the variation. Upon examination of these four features **(Table 2.3)** it was observed that they were the only features that had a high variance and high mean across the fingerprints; thus they could be highly discriminatory. However none of them showed any specificity to a particular species. Furthermore, the loading plots obtained also from the PCA analysis classified the *Salvia* species in the same three clusters as the hierarchical cluster dendrogram (Appendix 5).

In order to find species-specific features, a second analysis based on the magnitude of the variance for the normalized mean signal intensity of each feature across the 15 genotypes was used. Ten species-specific features were identified which were not previously detected by PCA since they had low mean across the fingerprints **(Table 2.3)**. These results imply that the PCA alone was not able to detect all the polymorphic sequences in the dataset since high variances were found for features with high and low mean signal intensity among the fingerprinted samples, and the PCA was only able to detect the features with high variance and high mean.

Furthermore, it is important to take into account that there was more than one feature that showed specificity to the same species, implying there were features with the same patterns of variation across the fifteen genotypes. Pearson bivariate correlation performed among the 14 features **(Appendix 6)**, indicated that there were positive significant correlations between H17 and J9 ($r = 0.98$, $P<0.01$), H17 and G4 ($r =0.99$, $P<0.01$), G13 and N7 ($r = 0.99$, $P<0.01$) and between N6 and I7 ($r = 0.83$, $P<0.01$). It is important to note that although these features were highly correlated may not necessarily imply that they possess high sequence similarity. However, J9, G4, N7 and I7 were eliminated from the set of polymorphic features since H17, G13 and N6 could explain most of the variation found in them. Based on the above analysis, only 10 features were selected since they were the most discriminatory and each had unique patterns of variation among the 15 genotypes.

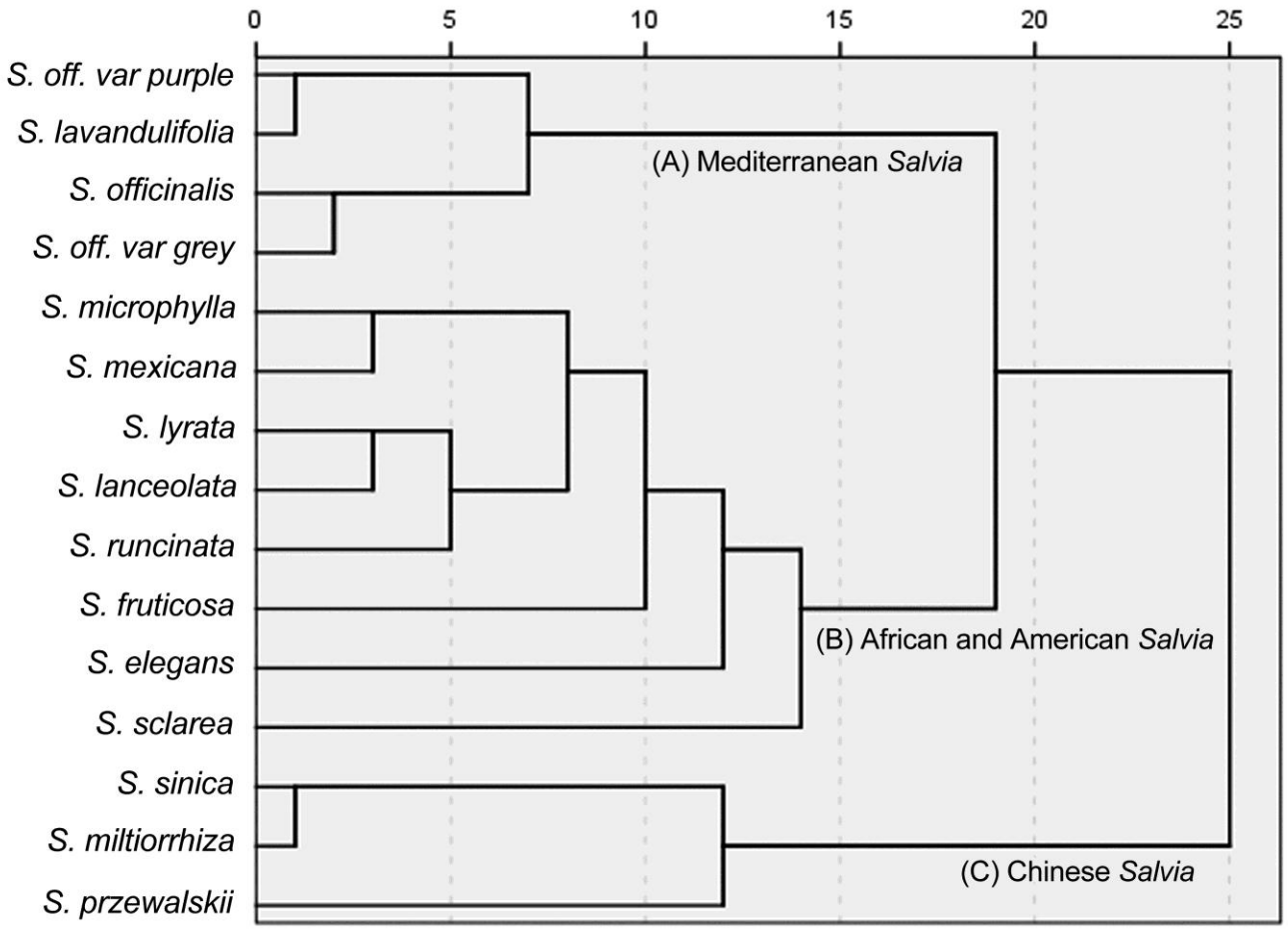


Figure 2.8. Dissimilarity dendrogram (Squared Euclidian distance, between groups linkage) for the SDA hybridization patterns of the fifteen genotypes using the 285 features. The steps of the dendrogram show the combined clusters and the values of the distance coefficients at each step; the values have been rescaled to numbers between 0 and 25, preserving the ratio of the distances between the steps.

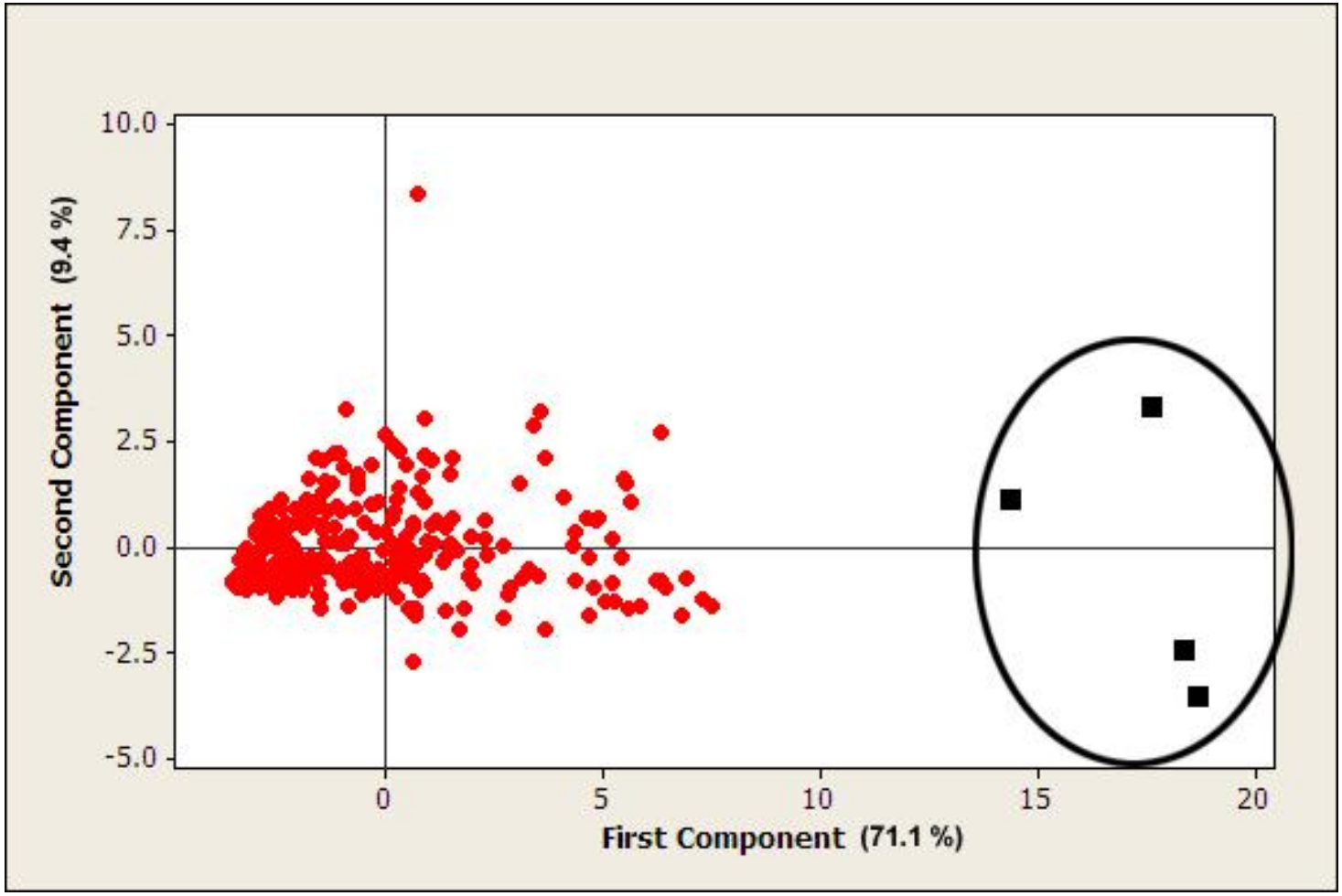


Figure 2.9. Principal component analysis plot for the 285 features. The first principal component accounts for 71.1% of variation and the second component explained only 9.4% of variation. The squares represent features that account for most of the variability found across the genotypes.

Table 2.3. Normalized mean signal intensities of the ten species-specific features and the 4 features chosen by PCA across the fifteen genotypes.

Genotypes	N12[b]	H17[b]	J9[b]	G4[b]	E13[b]	O1[b]	F5[b]	G13[b]	N13[b]	N7[b]	N6[a]	I7[a]	P4[a]	A16[a]
S. runcinata	5.71	3.60	10.91	14.52	27.72	21.41	40.01	0.76	155.16	6.42	376.02	289.39	198.08	263.41
S. microphylla	7.02	6.03	11.03	15.17	18.46	16.21	29.91	4.75	20.19	7.97	262.91	371.71	195.89	264.14
S. lanceolata	11.34	5.96	11.62	16.25	26.12	21.93	25.45	4.63	51.12	7.03	305.48	286.10	203.04	241.11
S. lyrata	3.02	3.29	12.35	18.50	24.71	17.28	49.75	0.80	23.58	6.99	308.68	343.08	154.35	259.88
S. sclarea	7.05	5.15	14.32	19.67	23.76	234.62	36.06	3.51	20.61	6.49	255.15	226.81	170.96	212.13
S. mexicana	7.71	8.04	15.03	16.04	19.40	15.64	23.67	6.43	33.80	7.89	254.40	258.83	295.14	201.49
S. fruticosa	13.02	13.36	17.63	21.87	19.64	20.88	46.03	10.26	23.03	14.46	273.78	361.46	165.68	189.39
S. elegans	13.97	15.56	18.55	25.63	17.89	28.98	37.36	145.77	35.82	135.89	198.30	199.62	223.37	218.84
S. sinica	196.56	3.07	24.55	20.68	13.90	30.72	61.12	0.56	16.80	3.48	123.33	97.94	137.93	178.12
S. przewalskii	163.45	5.64	33.78	26.32	38.07	30.98	201.59	3.48	12.60	5.39	119.45	113.04	194.76	241.75
S. miltiorrhiza	176.81	6.18	37.39	23.76	18.01	45.53	50.62	3.18	27.43	5.06	88.88	82.43	116.23	191.89
S. off. var purple	12.94	106.64	135.83	121.58	25.50	22.03	43.52	4.22	24.49	7.92	220.87	205.55	139.70	182.62
S. lavandulifolia	17.77	152.28	149.75	142.42	37.89	20.08	34.99	3.11	16.66	6.69	209.45	179.84	141.70	188.85
S. off. var grey	17.24	168.15	152.83	164.31	147.81	15.79	25.41	0.83	12.48	4.64	165.37	133.34	186.61	158.34
S. officinalis	15.61	137.17	162.70	152.96	217.07	13.91	22.58	2.53	15.65	4.03	131.90	209.33	165.74	191.65
Mean	**44.61**	**42.68**	**53.88**	**53.31**	**45.06**	**37.07**	**48.54**	**12.99**	**32.63**	**15.36**	**219.60**	**223.90**	**179.28**	**212.24**
Variance	**4891.02**	**3930.06**	**3707.57**	**3381.38**	**3329.22**	**3054.84**	**1918.97**	**1355.75**	**1252.42**	**1118.36**	**6760.21**	**8843.56**	**1898.37**	**1164.26**

[a] Features that were chosen by PCA.

[b] Species-specific features

Among the subset of the 10 features, species-specific features were found for *S. officinalis, S. sclarea, S. przewalskii* Maxim., *S. elegans* Vahl and *S. runcinata* L.f. **(Table 2.4)**. In addition, specific features which discriminated between species were found; for instance, feature N12 was found to be specific for *S. sinica, S. miltiorrhiza* and *S. przewalskii* and H17 may differentiate *S. officinalis* and *S. lavandulifolia* Vahl from the other species.

A second hierarchical cluster was performed using the 10 features selected above (3 chosen by PCA and 7 species-specific) **(Figure 2.10)**. It was found that the clustering of the species was consistent with the major clusters obtained with the full data set **(Figure 2.8)**. The differences observed within these two dendrograms were in Cluster B where five species were displaced relative to the original dendrogram. In the second dendrogram *S. microphylla* clustered with *S. lanceolata* and *S. lyrata* L. while *S. mexicana* L. *and S. runcinata* appeared to be more distantly related. In contrast, in the original dendrogram *S. microphylla* and *S. mexicana* clustered together while *S. runcinata, S. lanceolata* and *S. lyrata* were closely related. Therefore, it can be inferred that these 10 features were the most discriminatory for the fingerprinting of these fifteen genotypes.

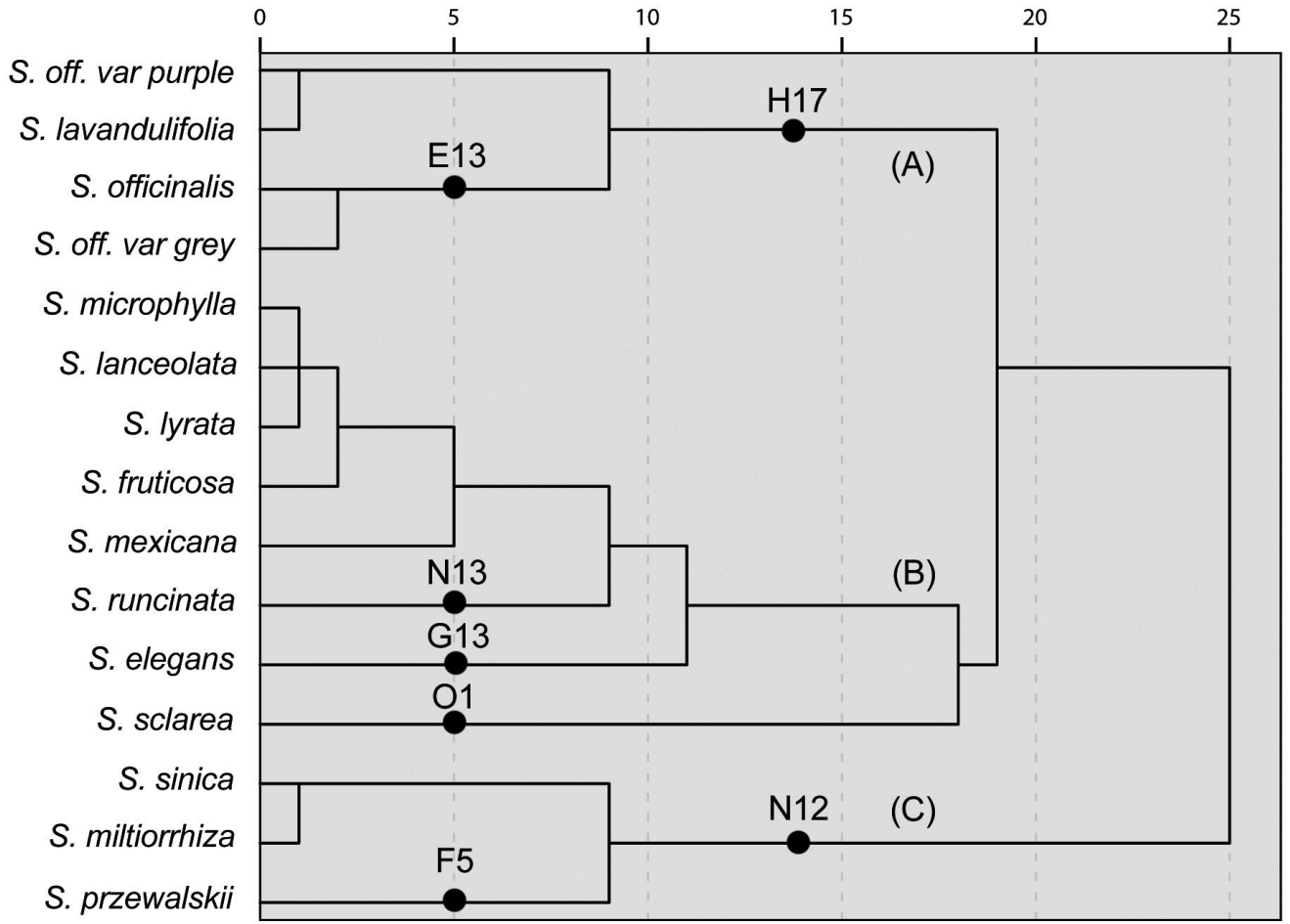


Figure 2.10. Dissimilarity dendrogram (Squared Euclidian distance, between groups linkage) for the SDA hybridization patterns of the fifteen genotypes using only the ten most discriminatory and species-specific features. The steps of the dendrogram show the combined clusters and the values of the distance coefficients at each step; the values have been rescaled to numbers between 0 and 25, preserving the ratio of the distances between the steps. The black shaded circles in the dendrogram represent species-specific features which discriminated across the species.

Table 2.4. Predicted locus/function of the 10 sequenced SDA features using blastN program through National Centre of Biotechnology Information (www.ncbi.nlm.nih.gov). Showing the best match as the putative identity for each sequence. E-value regarded as significant if < 1e-10. NA indicates the absence of significant data.

Feature ID	Length (bp)	Matching database entry	Putative identity	E Value	Specific to target
A16[a]	556	-	No hits	NA	
E13[b]	373	DQ673256.1	*Forsythia europaea psaA-psbB* fragment, chloroplast.	7e-151	Differentiate accessions of *S. officinalis*
		GQ996975.1	*Antirrhinum majus* cpl protease proteolytic subunit protein (*cplP*) gene, complete cds, chloroplast	8e-150	
F5[b]	218	DY322087.1	*Ocimum basilicum* uncharacterized cDNA sequence	6e-14	*S. przewalskii*
G13[b]	145	-	No hits	NA	*S. elegans*
H17[b]	612	-	No hits	NA	*S. officinalis* *S. lavandulifolia*
N12[b]	423	FJ148301.1	*Daucus carota* subsp. *sativus* uncharacterized sequence	2e-06	*S. miltiorrhiza* *S. sinica* *S. przewalskii*
N13[b]	556	-	No hits	NA	*S. runcinata*
N6[a]	250	-	No hits	NA	
O1[b]	118	-	No hits	NA	*S. sclarea*
P4[a]	526	DQ673256.1	*Forsythia europaea psaA-psbB* fragment, chloroplast.	2e-77	
		DQ983917.1	Jacobaea uniflora isolate SGO10 tRNA-Met (*trnM*) gene, partial sequence; ATP synthase epsilon subunit (*atpE*) and ATP synthase beta subunit (*atpB*) genes, complete sequence; and ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit (*rbcL*) gene, partial sequence; chloroplast	2e-65	

[a] Features that were chosen by PCA.

[b] Features that are part of the seven species-specific features.

2.3.3 The sequence identity of the most discriminatory and species-specific features

The species-specific features together with the three features chosen by PCA were sequenced and analyzed using blastN and blastX (www.ncbi.nlm.nih.gov) **(Table 2.4 and Appendix 7)**. Four out of the ten features matched with homologous sequences in GenBank. Features E13 and P4 corresponded to known chloroplast loci. Feature N12 had a good match to an uncharacterized genomic DNA, and F5 corresponded to an uncharacterized gene and the other six features were not recognized as known DNA sequences or proteins.

Furthermore, the efficiency of the suppression PCR effect may be estimated by looking at the sequences of these features **(Appendix 7)**. Theoretically, only the sequences with different adaptors at their ends should have been exponentially amplified during the suppression PCR **(Figure 1.5)**. It was found that seven of the features had sequences with both adaptors at their ends, indicating they may be tester specific since they were found negative for the driver target. The sequences of features P4 and N6 were of poor quality at the begging and end of the reading. Although, both fragments were re-sequenced, it was not possible to recognize both adaptors in each sequence. It is possible that the poor quality of the reading may be attributed to the low quality of the DNA fragments. Finally, the sequence of feature F5 was found to have the same adaptor sequences at both ends (Adaptor 2) which implies that the suppression PCR effect may not have been fully efficient, although this feature was found also negative for the driver target. The reasons for this will be explained further in the discussion.

2.4 DISCUSSION

2.4.1 Subtraction efficiency

The subtraction technique was able to eliminate about 88% of common DNA sequences between the tester (*Salvia*) and driver pools. This result is comparable with the percentage of positive subtracted clones found between *Dendrobium* species (85%-76%) (Li et al., 2006). However, the subtraction efficiency is lower than the one obtained for the prototype SDA for angiosperms (Jayasinghe et al., 2007) where 3% of the features were found to be positive for the driver (non-angiosperm) target. Therefore, it may be concluded that although the subtraction was efficient in eliminating the common sequences, the subtraction protocol could be optimized further to obtain a higher efficiency. Although, the root cause of this reduce efficiency was unclear, it may be possible that deficiencies during the digestion, ligation, hybridization or PCR amplification could have affected. As digestion and ligation were positively verified during the subtraction process (**Section 2.2.2.2**); errors or deficiencies during the suppression PCR or hybridization process may have been the cause.

The presence of the same adaptor at both sites of the sequences of feature F5 (**Section 2.3.3**), may be a proof that the suppression PCR did not work efficiently, since it allowed the amplification of type b molecules. Type b molecules **(Figure 1.5)**, which contain complementary sequences on the ends, should have been suppressed during the primer annealing step where the hybridization kinetics favors the formation of "panhandle-like" structures, which prevents the primer annealing and further extension (Diatchenko et al., 1996). Therefore, if some of these type b molecules were

exponentially amplified during subtraction, then the presence of these molecules in the array could have lowered the subtraction efficiency. During the subtraction, the products of first and second PCR were analyzed by gel and the results were similar to what it is described in the user manual, however it seems that the PCR conditions could be optimized in future experiments to avoid amplification of type b molecules.

Furthermore, the hybridization parameters used could have also affected the subtraction efficiency. For example, the concentration of driver used may have not been enough to remove all similar sequences. Thirty-fold excess of driver was used for subtraction as recommended by the kit manufacturer (Clontech). However, even higher concentrations of driver may have been employed, as theoretically higher concentrations of driver will subtract the sequences that are partially homologous between the tester and driver enriching for those highly specific *Salvia* sequences (Diatchenko et al., 1996). Moreover, the subtraction hybridization temperature of 68^0C which was also recommended by the kit manufacturer could have been too high to be able to eliminate the sequences that are partially homologous between the pools. At a high subtraction hybridization temperature only the highly similar sequences between the driver and tester would hybridize and be subtracted, leaving the less similar sequences single-stranded to be amplified by PCR (Gadgil et al., 2002). Therefore, the temperature used in the two rounds of subtraction hybridization could be performed at lower temperatures, in order to eliminate the sequences that are partially homologous between the pools. Although, it is not possible to determine that the above conditions were in fact the exact cause of the inefficient elimination of common sequences in this

experiment, the modification of these parameters in future DNA subtractions could clarify the effect of these conditions on the subtraction efficiency.

2.4.2 Scoring of the microarray

The raw signal intensity was used to perform the scoring in this present study since the binomial data had low reproducibility (as explained in **Section 2.2.5.2**). A study based on DArT found that the biallelic assessment limited the range of possible analysis in 17 *Eucalyptus* individuals (Lezar et al., 2004). Furthermore, another DArT study on *Arabidopsis thaliana* successfully identified segregating markers in a F_2 population based on the differences in the intensity of the hybridization signal (Wittenberg et al., 2005). One common feature between the current study and the previous *Eucalyptus* DArT study is that the species analyzed are mainly cross-pollinated (Claßen-Bockhoff, 2007; Pound et al., 2002). Therefore, the cross-pollination in these species could have increased the level of heterozygosity, which will generate intermediate signal intensities that could have complicated the analysis of binomial data. Consequently, the use the direct comparison between signal intensities is the preferred option for a cross-pollinated genus such as *Salvia*, since the assessment based on presence and absence (dominant) could have potentially misleading effects due to the higher heterozygosity expected in this genus.

2.4.3 Fingerprinting of fifteen *Salvia* genotypes

The hierarchical clustering performed demonstrated the ability of the SDA to fingerprint closely related species and accessions within species. For instance, it was possible to

differentiate among three accessions of *Salvia officinalis* and their related species *S. fruticosa* and *S. lavandulifolia*. These three species are commonly misidentified due to their morphological similarity (Reales et al., 2004), which has lead to substitution in commercial products. In the present study, S. *lavandulifolia* was found to be closely related to three accessions of *S. officinalis*, in particular to *S. officinalis* var. purple, while *S. fruticosa* was found in another separate cluster from these two species **(Figure 2.8 and 2.10)**. These results are in agreement with the taxonomical classification made by Reales et al. (2004), where S. *lavandulifolia* was found to be a subspecies of *S. officinalis* and *S. fruticosa* was clearly differentiated from these two taxa. Furthermore, SDA could also effectively differentiate among related species of *S. miltiorrhiza* (*S. przewalskii* and *S. sinica*) which are commonly used as Danshen, although chemical fingerprinting by HPLC have found that these species do not meet with the required standards of Chinese Pharmacopoeia (Cao et al., 2008; Li, 2008). Based on the above results it is possible to conclude that this *Salvia* specific SDA could be a useful tool for authentication purposes of *S. officinalis*, *S. fruticosa* and *S. miltiorrhiza*.

Moreover, SDA was capable of fingerprinting species that were not used in its construction. For instance, it was possible to fingerprint *S. lanceolata, S. microphylla* and *S. fruticosa*, which grouped with other *Salvia* species of similar geographical origin in the hierarchical cluster performed with the 285 features **(Figure 2.8)**, even though they were not part of the original subtraction pool. Similar results were found by Jayasinghe et al. (2009), who developed an angiosperm specific SDA that was able to fingerprint species outside the initial angiosperm pool. Therefore, this SDA may have a wider applicability in fingerprinting other species of *Salvia* apart from the ones used to

construct the array, with the advantage that no prior knowledge of sequence information is needed and it is not necessary to screen for primers which amplify for polymorphic loci.

2.4.4 Diversity analysis

The hierarchical analysis revealed genetic relationships consistent with geographical origins **(Figure 2.8 and 2.10)**. The dendrogram shows three major clusters. The first cluster included the European *S. officinalis* and *S. lavandulifolia*, the second cluster had the native American and African species together with *S. sclarea* and *S. fruticosa*, and the third clustered the Chinese *Salvia*. However, these results contradicted previous phylogenetic studies performed in the tribe Mentheae and in the genus *Salvia* using the sequences of the amplified nuclear rDNA ITS and chloroplast DNA regions of *rbcL*, *trnL-F* and *psbA-trnH* (Walker and Sytsma, 2007; Walker et al., 2004), where three lineages of *Salvia* were postulated. The first clade [*Salvia* Clade I sensu Walker and Sytsma (2007)] is represented by European, Central African, southern African and west Asian species. *Salvia* Clade II contains *Salvia* section Audibertia which is restricted to the Californian Floristic Province and adjacent deserts, and *Salvia* subgenus Calosphace which occurs primarily in Central and South America. Finally, the *Salvia* Clade III includes Asian, Northern African and Mediterranean species.

There are some clear differences between the Walker and Sytsma's classification (2007) and the one obtained with the SDA. For example, in the present study the Mediterranean *Salvia* (Cluster A), together with *S. sclarea* (Cluster B) **(Figure 2.8)** are part of Salvia Clade I sensu Walker and Sytsma (2007). Also, most of the African *Salvia* species are

in a different clade from American species according to Walker and Sytsma (2007). The only concordance between the two classifications is the clustering of *S. sinica*, *S. miltiorrhiza* and *S. przewalskii*, which are found into Cluster C in the preset study and in *Salvia* Clade III sensu Walker and Sytsma (2007). Two factors may account for the differences: Firstly, the work of Walker and Sytsma (2007) and Walker et al. (2004) was based on different nuclear (rDNA ITS) and chloroplast regions from the ones that were identified as the more discriminatory sequences in this SDA work. Therefore, different regions could have different polymorphisms which may possibly give different distances among the species. Secondly, during the construction of the SDA, the *Salvia* pool was enriched with *S. miltiorrhiza*, *S. sinica* and *S. officinalis*. Therefore, the SDA could be overrepresented with sequences from these three species, and as a result the phylogenetic analyses obtained from this array could be biased on the distances given across the species and major clusters.

It is important to note that the main aim of this study was not to perform a phylogenetic analysis; however, the results showed genetic relationships consistent with geographical origins which may imply that SDA could be useful for phylogenetic analysis. A previous study have shown the utility of the SDA for inferring genetic relationships consistent with the Angiosperm Phylogeny Group (2009) classification (Bremer et al., 2009). For instance, the angiosperms specific SDA has shown to be useful in classifying different families with the respective clade and species within their correspondent families (Jayasinghe et al., 2009). However, in order to apply this technique for phylogenetic analyses in *Salvia,* a more comprehensive array would have to be constructed with equal genomic representations from all *Salvia* subgenera, and a wider

range of species would have to be genotyped in order to obtain a more detailed phylogenetic and evolutionary analysis of the genus.

2.4.5 Identity of the most discriminatory and species- specific features

Among the most discriminatory and species-specific features, two were identified as chloroplast loci **(Table 2.4)**. The other unknown and uncharacterized sequences are almost certainly part of nuclear DNA, since if they were part of chloroplast or mitochondrial DNA they would have shown a match in GenBank (which has an extensive amount of chloroplast and mitochondrial database entries).

The presence of these polymorphic nuclear DNA features suggests that the SDA may be a more reliable approach for fingerprinting closely related species and hybrids. The reason for this may be that the detection of hybridization/introgression is not reliably accomplished by examination of chloroplast DNA since it is uniparentally inherited (Fazekas et al., 2009). In addition, dependence on a single nuclear locus could have misleading results since the hybrid could be homozygous at many loci (Chase et al., 2005). Future studies could evaluate the potential of the SDA to identify hybrids in *Salvia*, for instance it will be of interest to fingerprint hybrids from *S. officinalis* with *S. fruticosa*, since these two species may produce hybrids spontaneously or from breeding programs which could be misidentified or commercialized as any of the parents (Dudai et al., 1999; Putievsky et al., 1990; Reales et al., 2004).

Furthermore, these species-specific features could be developed as PCR-based markers in order to obtain a fast and easy fingerprint of the species. For instance, it would not be

advisable to use the entire array to fingerprint only one or a small number of species since it would be more time and consuming expensive than PCR. Instead, specific primers could be developed for the amplification of these sequences in order to find if there are variations in amplification between closely related species that could lead to the unequivocal identification of the species of interest.

Finally, it is important to note that although these 10 features were found to be highly polymorphic for the 15 genotypes analyzed, it does not mean that the other features in the array are not useful. For instance, some of these other features could be highly polymorphic for other species or accessions of *Salvia* or instead they could be monomorphic across different *Salvia* which could imply they could be *Salvia* specific.

2.5. CONCLUSIONS

To the best of authors' knowledge, this is the first fingerprinting array constructed for *Salvia* species. Using this array it was possible to fingerprint 15 *Salvia* genotypes and to construct a hierarchical cluster which was found to be consistent with the geographical origin and was able to differentiate closely related species of *S. officinalis* and *S. miltiorrhiza*. Most importantly, SDA has shown to have potential advantages over other fingerprinting techniques:

(i) The subtraction technique made it possible to enrich the SDA with a set of unique sequences for the taxa under study, which opens the possibility to

fingerprint species of *Salvia* that were not used to construct the array, with the advantage that no prior knowledge of sequence information is needed and it is not necessary to screen for primers which amplify for polymorphic loci.

(ii) its ability to fingerprint based on polymorphic regions found on chloroplast and nuclear DNA increases the possibilities of being able to differentiate closely related species and hybrids.

(iii) SDA has shown to be a powerful technique, able to screen for species-specific features which could be potential PCR based markers used for the authentication of species.

Based on the above results it is possible to conclude that SDA is a technique that can effectively isolate highly variable DNA sequences specific to a genus as large as *Salvia* without preliminary sequence information. Therefore, SDA is a potential technique to fingerprint non-model plants where few markers are available.

CHAPTER 3

Fingerprinting of geographical populations of *Salvia miltiorrhiza* using the *Salvia*-SDA

3.1 INTRODUCTION

This chapter describes the use of SDA to fingerprint *S. miltiorrhiza* populations selected from different primary production areas in China. The data obtained was used for assess the genetic diversity and to screen the SDA for highly polymorphic sequences that could differentiate between geographical populations of *S. miltiorrhiza*. Finally, the SDA genetic profiles were correlated with chemical and morphological profiles obtained from previous studies.

Recently, the rapid growth of Traditional Chinese Medicine (TCM) has raised issues about its quality and safety (Leung and Cheng, 2008). In order to ensure the efficacy and safety of the herbs used in TCM a reliable approach for quality control is needed. Chromatographic fingerprinting methodology such as HPLC has been regarded as a useful method for the quality control of herbal medicines since it can separate the complex composition of the samples into subcomponents that are representative of the chemical profile (Chen et al., 2009b). Danshen, the dried root of *S. miltiorrhiza* and its products (pharmaceutical preparations) have been extensively studied by chromatographic techniques which can identify and asses the quality by separation and identification of the major bioactive compounds (Liu et al., 2007; Yang et al., 2006a; Zhou et al., 2006).

The major disadvantage of using chromatographic fingerprinting in *S. miltiorrhiza* roots is that the content of the bioactive compounds varies significantly depending on the environmental and agricultural conditions in which it is grown (harvest time, climatic and soil conditions), even geographical origin has been found to be an important factor (He et al., 2010; Li et al., 2009b). According to previous studies on genetic diversity of *S. miltiorrhiza*, the genetic variance existing within populations of plants cultivated in the same location is higher, than among populations (Guo et al., 2002; Song et al., 2010). Therefore, if consistent content of bioactive compounds is needed in order to improve Danshen quality, not only environmental and agricultural conditions should be monitored but also it would be important to identify cultivars which combine good agricultural traits and could produce high yield of bioactive compounds. However, few cultivars have been developed despite the long history (A.D. 102-200) of therapeutic application in China (Li, 2008; Song et al., 2010). Consequently, optimal populations or individuals plants should be identified which could serve as parental lines in future breeding programs.

Molecular markers linked to desirable agricultural traits and high yield of bioactive compounds will be useful to screen for optimal populations or individuals in future breeding programs for *S. miltiorrhiza*. Different DNA fingerprinting techniques have proved useful in genotyping different *S. miltiorrhiza* samples and other closely related *Salvia* species. Previous studies include RAPD (Guo et al., 2002), ITS (Han et al., 2010; Xu et al., 2009), inter-simple sequence repeat (ISSR) (Song et al., 2010), Amplified length polymorphism (AFLP) (Wang et al., 2007), Sequence related amplified

polymorphism (SRAP) (Song et al., 2010), Conserved region amplification polymorphism (CoRAP) (Wang et al., 2009) and Simple sequence repeats markers (SSR) derived from Expressed Sequence Tags (EST) (Deng et al., 2009). To date, only the studies performed by Han et al. (2010) and Xu et al. (2009) have analyzed the relation between the molecular profiles obtained with ITS sequences and HPLC profiles of important bioactive components; however no significant relationship has been found. Therefore, there is a lack of molecular markers linked to desirable agricultural traits in *S. miltiorrhiza* that could be used for fingerprinting and for breeding purposes.

Studies carried out during 2005-2007 at the RMIT Health Innovations Research Institute performed agronomical and chemical analyses in seven Australian-grown Danshen populations, in order to identify the optimal environmental and agricultural conditions for cultivation; and also to identify the optimal Danshen-population which combines good root yield and high content of bioactive compounds (Li et al., 2009a). Significant differences were found in total weight of fresh roots-per plant, content of tanshinones and salvianolic acid B among the seven populations. Their results showed that there are significant differences in the content of biomarker compounds across populations even when they were cultivated under the same conditions, which imply that the selection of a specific Danshen-population is an important factor if higher content of bioactive compounds want to be achieved. Therefore, it will be of interest to assess the genetic variability of these populations and to correlate their molecular fingerprints to their chemical profile in order to identify potential molecular markers associated to agricultural traits and production of bioactive compounds which could

assist in the screening of optimal parents from which future breeding programs could be developed.

The objectives of the experiments described in this chapter are: (1) to evaluate the potential of SDA to fingerprint geographical populations of *S. miltiorrhiza* (2) to assess the genetic diversity among populations of *S. miltiorrhiza* (3) to establish if the differences found in content of bioactive compounds and other agricultural traits previously studied can be related to the genetic profiles.

3.2 MATERIALS AND METHODS

3.2.1 Plant material

Salvia miltiorrhiza and *S. sinica* plants were obtained from seeds which were used in a previous study (Li et al., 2009a). These seeds were collected from different provinces in China. Eight lines were employed in the previous study; however due to seed availability only six lines were used for this study (*S. miltiorrhiza f. alba* (Shandong), *S. miltiorrhiza* Shandong, Shanxi, Henan, Hebei and *S. sinica* (Zhejiang province) **(Table 2.1)**. Seeds for each of the lines were germinated in Petri dishes and transplanted in 15-cm diameter pots. The potting mix was prepared by mixing 50% sand and 50 % of a general-purpose potting mix (Yates^{TM}, NSW), which was used to fill about 40% of each pot; then only the general-purpose potting mix was used to filled up each pot. Five seedlings from each line of *S. miltiorrhiza* and *S. sinica* were transplanted in the pots which were kept in glasshouse under controlled temperature 20± 3^0C.

3.2.2 Fingerprinting of geographical populations of *S. miltiorrhiza*

3.2.2.1 DNA extraction and labeling of Target DNA

Genomic DNA was extracted from individual plants as described in **Section 2.2.2.1.** The DNA of the five seedlings was then pooled in order to obtain representations for each of the *S. miltiorrhiza* and *S. sinica* populations. Each pool had equal amounts of genomic DNA per plant (about 200 ng/plant) to obtain approximately 1 µg of DNA for each representation. Subsequently, each DNA pool was digested overnight with *AluI* and *Hae*III (Fermentas) and column-purified (Qiagen^{TM} QIAquick PCR Purification Kit, Qiagen). The concentration and purity of the DNA pools were evaluated spectrophotometrically. Approximately 150 ng of purified digested DNA was labeled with Biotin-11-dUTP using the Biotin DecaLabel^{TM} DNA Labeling Kit (Fermentas, ON, Canada).

3.2.2.2 SDA hybridization

The same microarray developed for fingerprinting *Salvia* species (Chapter 2) was used for all the subsequent studies described in this chapter.

During the construction of the SDA, the tester pool was enriched with *S. miltiorrhiza* and *S. sinica* **(Table 2.1)**, consequently a high number of the subtracted fragments printed on the array may be specific to these two species. As a result, most of the features displayed a strong hybridization signal after hybridization of *S. miltiorrhiza* and *S. sinica* targets using the protocol described in **Section 2.2.3.2.** Therefore, the

following modifications were made to labeling and hybridization protocol to increase the stringency of the array and achieve higher discrimination among populations.

- Firstly, the biotin-labeled sample was reduced from 30 ng employed previously to only 20 ng. This change did not affect the reproducibility (>0.91).
- Secondly, the temperatures and times of the four stringency washes were modified: The slides were washed twice in 1 x SSC, 0.1% SDS at 45^0C for 8 min, twice in 0.1 x SSC, 0.1% SDS at 45^0C for 10 min and once in 0.1 x SSC at room temperature for 5 min.

Then, detection of the biotinylated DNA targets bound on the array was performed by the protocol modified from Mirus Label IT® µArray® Biotin Labeling Kit (Mantri et al., 2011) described in **Section 2.2.3.2**

3.2.3 Analysis of the *Salvia*-Array

3.2.3.1 Scanning, quantification and data analysis

Scanning of the slides, quantification and subsequent analysis (background correction, flagging, normalization and mean of the signal intensity across technical and biological replicates) were performed as described for the fingerprinting and data analysis of *Salvia* species **(Section 2.2.5)**. The normalized mean signal intensity was then used for all subsequent statistical analyses.

3.2.3.2 Statistical analysis

The normalized mean data was used to perform PCA, analysis of variance, Pearson bivariate correlation and hierarchical cluster analysis as described in **Section 2.2.5.3**.

Data obtained from previous agronomical and chemical analyses on the same geographical populations of *S. miltiorrhiza* and *S. sinica* (Sheng, 2007) was used for correlation analyses. Each of the following parameters: number of side branches, aerial weight, number of roots, maximum root diameter, root weight and the content of bioactive constituents, was correlated with the normalized mean signal of the most discriminatory features by performing Pearson bivariate correlations (SPSS version 17.0) and regression analysis (Microsoft Excel). The data of the parameters obtained from the previous study can be found in **Appendix 8**.

3.2.4 Sequencing of selected features

Amplification products of the most discriminatory features were sequenced and nucleic acid and protein homology searches were performed using blastN and blastX programs through the National Center of Biotechnology Information (www.ncbi.nlm.nih.gov) as described in **Section 2.2.6.**

3.3 RESULTS

3.3.1 Fingerprinting of geographical populations of *S. miltiorrhiza* and *S. sinica*

The hierarchical cluster analysis constructed with the signal intensity of 285 features provided a clear differentiation between *S. sinica* and the five populations of *S. miltiorrhiza* **(Figure 3.1)**. Among the populations of *S. miltiorrhiza*, the ones from Shandong and Hebei province clustered together and were closely related to the population from Henan. The other two populations, *S. miltiorrhiza f. alba* (Shandong province) and the one from Shanxi province were found more distantly related.

Principal component analysis performed with the signal of the full set of features (total of 285) indicated that the first principal component accounted for 94.6% of the variation and the second component explained only 2.5% of the variation. Together, the first and second components accounted for 97.1% of the total variability of the data **(Figure 3.2)**. This analysis was able to associate the features with similar patterns of variation. For instance, features that clustered close to zero had low variance among the populations while the features that were distributed throughout the plot were features that presented the highest variances **(Table 3.1)**. Therefore, the twelve features that were more distant from zero on the X axis were chosen since they could be highly discriminatory. Among these 12 features, two of them (O14 and L5) were found to be features that gave high signal intensities for the driver target, which implied they were not *Salvia* specific. Consequently, these two features were excluded from further analyses. Furthermore, the loading plots obtained also from the PCA analysis were found to separate out clearly the populations from Henan and Hebei province form the population of Shandong. This

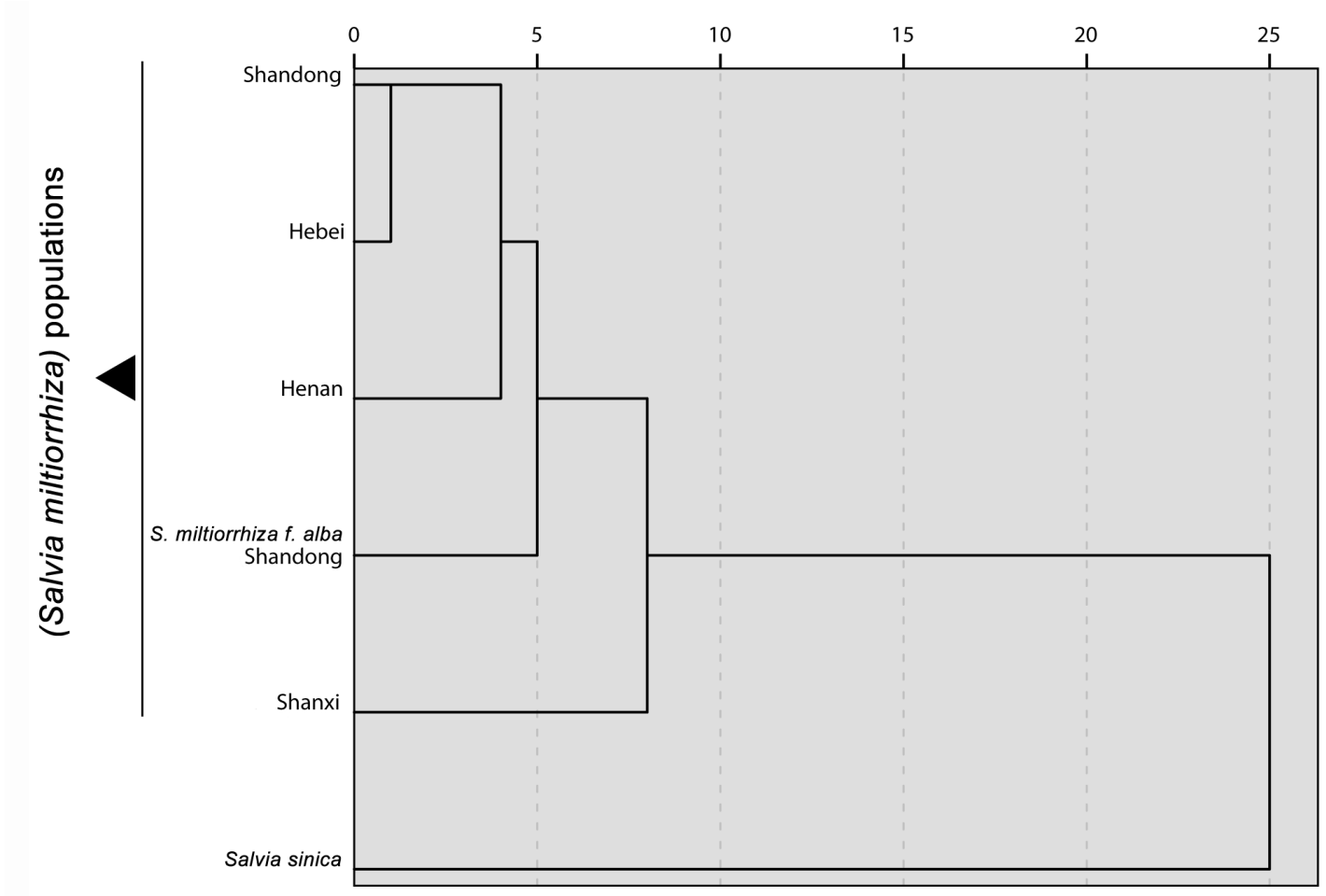


Figure 3.1. Dissimilarity dendrogram (Squared Euclidian distance, between groups linkage) for the SDA hybridization patterns of the five lines of *S. miltiorrhiza* and one of *S. sinica* using 285 features. The steps of the dendrogram show the combined clusters and the values of the distance coefficients at each step; the values have been rescaled to numbers between 0 and 25, preserving the ratio of the distances between the steps.

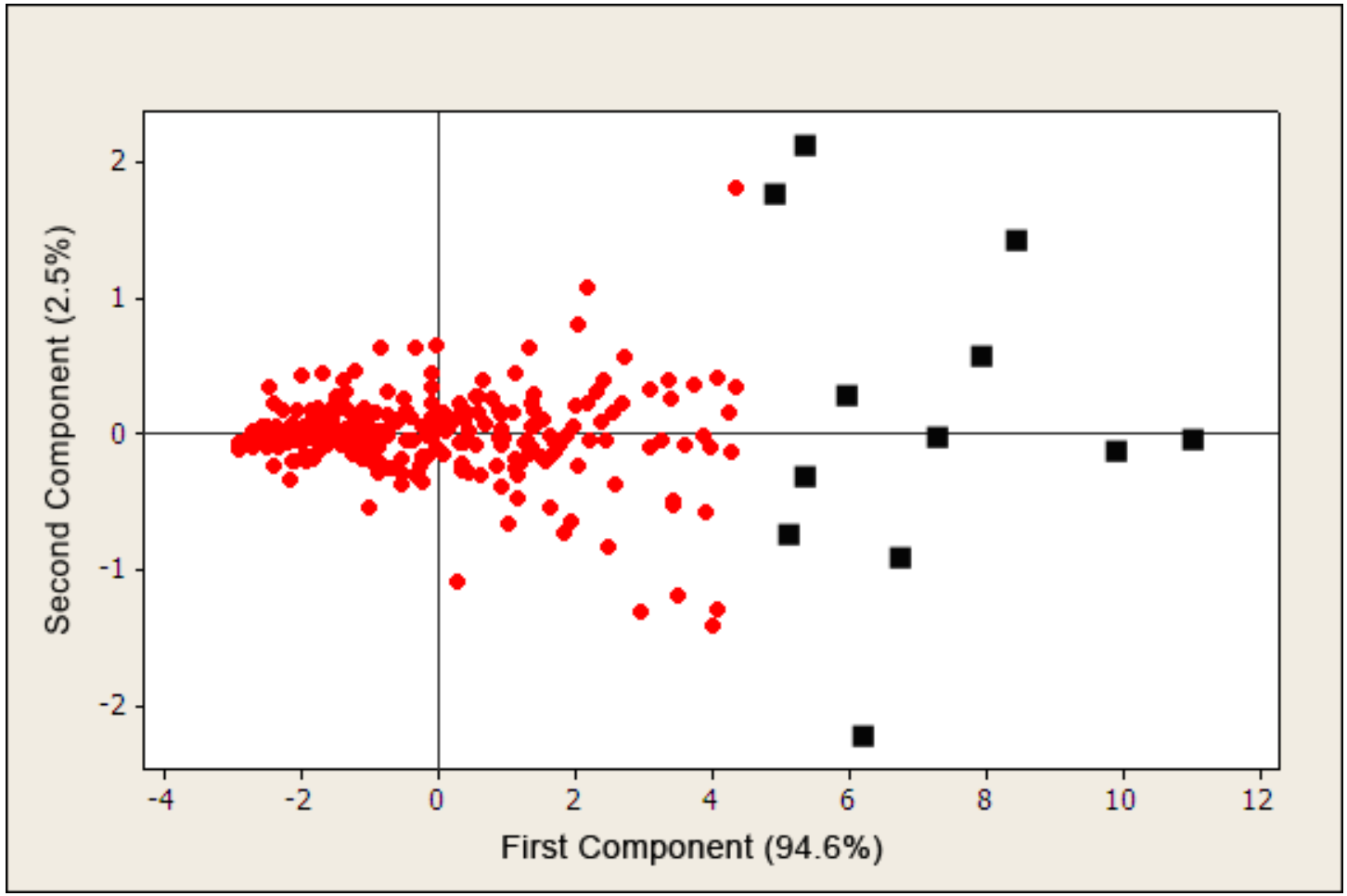


Figure 3.2. Principal component analysis plot for the 285 features. The first principal component accounts for 94.6% of variation and the second component explained only 2.5% of variation. The squared shaped spots represent the 12 most informative features.

Table 3.1. Normalized mean signal intensities of the 12 features chosen by PCA across the five lines of *S. miltiorrhiza* and one of *S. sinica*.

Feature / Line	H17	I5	C11	A5	G9	B7	A11	K2	K6	I7	O14*	L5*
S. sinica	**163.81**	**152.33**	**53.35**	**144.30**	**88.45**	**110.34**	**36.84**	**53.37**	**89.27**	**77.17**	**121.61**	**169.77**
Shandong province	**142.16**	**184.49**	**114.62**	**129.45**	**103.04**	**111.98**	**86.91**	**104.32**	**131.78**	**106.05**	**138.97**	**137.36**
Shanxi province	**183.51**	**139.29**	**109.79**	**132.07**	**103.36**	**118.99**	**120.84**	**70.36**	**126.88**	**96.11**	**121.09**	**143.88**
S. miltiorrhiza f. alba (Shandong)	**196.29**	**140.85**	**137.84**	**141.72**	**146.18**	**124.98**	**90.81**	**106.11**	**114.98**	**120.25**	**120.59**	**157.62**
Hebei province	**174.27**	**165.07**	**149.99**	**139.66**	**76.17**	**106.87**	**84.23**	**87.69**	**135.72**	**119.63**	**156.53**	**127.61**
Henan province	**192.74**	**185.38**	**128.22**	**130.43**	**111.64**	**99.96**	**104.74**	**109.65**	**133.14**	**90.41**	**111.11**	**122.09**
Mean	**175.46**	**161.23**	**115.64**	**136.27**	**104.81**	**112.18**	**87.39**	**88.58**	**121.96**	**101.60**	**128.32**	**143.06**
Variance	**408.76**	**422.95**	**1149.4**	**40.80**	**570.31**	**78.19**	**799.40**	**514.32**	**310.49**	**288.91**	**272.54**	**327.42**

* These features were found to be positive for the driver target and therefore were excluded from further analyses

differentiation was not found in the dendrogram since the populations from Shandong and Hebei province clustered together (Appendix 5).

In order to detect all the useful polymorphic features, the magnitude of the variance for the full set of features was also examined. It was found that the two features with the highest variances (C11 and A11), which were *S. miltiorrhiza* specific **(Table 3.1)**, were also detected by PCA. The data above indicated that the PCA alone was able to detect all the polymorphic sequences in the dataset since high variances were found only for features with high mean signal intensity. These results were in contrast to the PCA analysis performed in Chapter 2, where all highly polymorphic sequences were not detected, may be due to the fact that features with high and low mean signal intensities had both high variances.

In addition, Pearson bivariate correlation was performed on the 10 features detected by PCA in order to refine the subset of the most discriminatory features. Feature C11 was found to be correlated to features K6 ($r = 0.82$, $P < 0.05$) and I7 ($r = 0.85$, $P < 0.05$) **(Appendix 9)**, indicating that C11 could explain most of the variation found in the other two features. Based on the above analysis, only a subset of eight features was determined to be as the most discriminatory.

Finally, a second hierarchical cluster analysis was performed **(Figure 3.3)** using the 8 features chosen above to determine their discriminatory power across the populations analyzed. A comparison between this new dendrogram with the original constructed with the full set of features **(Figure 3.2)** revealed that the genetic distances between the

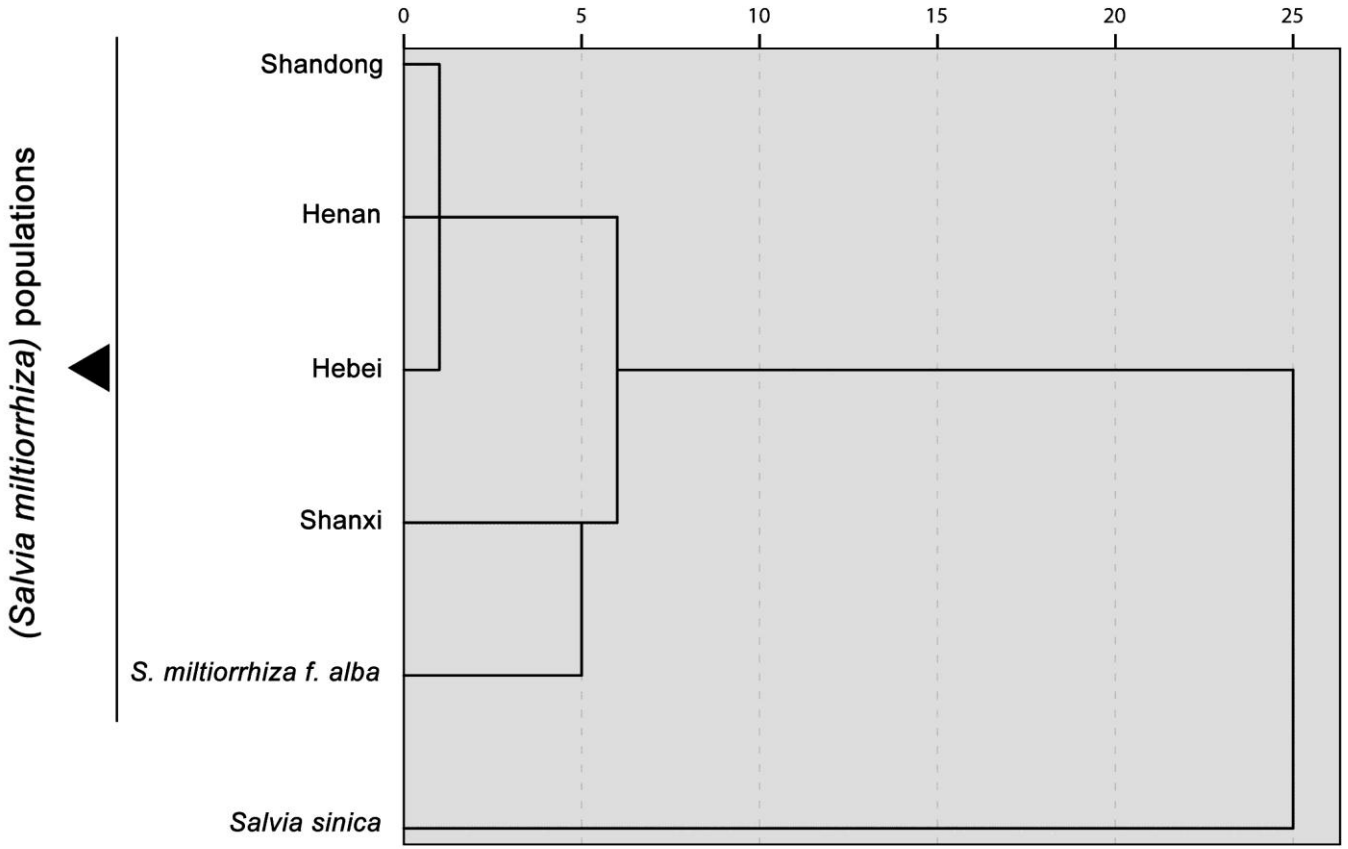


Figure 3.3. Dissimilarity dendrogram (Squared Euclidian distance, between groups linkage) for the SDA hybridization patterns of the five lines of *S. miltiorrhiza* and one of *S. sinica* using only the 8 most discriminatory features. The steps of the dendrogram show the combined clusters and the values of the distance coefficients at each step; the values have been rescaled to numbers between 0 and 25, preserving the ratio of the distances between the steps.

provinces of Henan, Shandong and Hebei province, and between *S. miltiorrhiza f. alba* (Shandong province) and Shanxi populations were reduced by comparison. In addition, both dendrograms were able to distinguish *S. sinica* as an out-group and were consistent in the relationships shown among the populations from Shandong and Hebei province. Consequently, it may be inferred that these eight features were able to adequately fingerprint the populations of *S. miltiorrhiza* and distinguish them from *S. sinica.*

3.3.2 The sequence identity of the interesting features

Amplification products of the eight most discriminatory features were sequenced together with the amplification products of O14 and L5 features. Sequences of only three features matched with known gene coding regions in GenBank **(Table 3.2)**. Features I5 and O14 significantly matched to chloroplast loci, I5 had a significant alignment with a *psaI* and *ycf4* fragment and O14 matched perfectly with a hypothetical chloroplast open reading frame 1 (*ycf1*). Feature L5 was found to be part of the 18S rRNA gene and the seven remaining features did not match to any database entry.

Furthermore, the efficiency of the suppression PCR effect may be estimated by observing at the sequences of these features since only the sequences with different adaptors at their ends should have been exponentially amplified during the suppression PCR. It was found that nine features had sequences with both adaptors at their ends **(Appendix 10)**; however out of this nine, the sequences of features O14 and L5 were found to be positive for the driver target. This result may indicate that there was not enough driver to remove these sequences during the hybridization steps in the subtraction. The reasons for this will be explained further in the discussion. The other

seven sequences that had both adaptors at their ends could be tester specific since they were found negative for the driver target. Finally, the sequence of feature A11 was found to have the same adaptor sequences at both ends (Adaptor 2), which implies that the suppression PCR effect may not have been fully efficient (as explained in **Section 2.4.1)**.

3.3.3 Correlation of the chemical / agronomical dataset with the molecular profile

The normalized signal intensities of the 8 most discriminatory features were correlated with the agronomical traits and the content of four major bioactive constituents of *S. miltiorrhiza* (cryptotanshinone, tanshinone I, tanshinone IIA and salvianolic acid B) **(Appendix 8)**.

Linear correlations were found between the signal strength of two features (C11 and K2) and the agronomical traits **(Table 3.3)**. The signal strength of C11 was found to be inversely correlated to aerial weight, root number and root weight. The signal strength of K2 was also inversely correlated to root number, maximum root diameter and root weight. Therefore, the signal strengths of both C11 and K2 correlated to the root weight and number of roots of the plant, while there were no correlations with the number of side branches.

Additionally, positive correlation was found between the signal strength of feature K2 and content of cryptotanshinone ($r = 0.98$; $P < 0.01$), tanshinone I ($r = 0.85$; $P < 0.05$) and tanshinone IIA ($r = 0.91$, $P < 0.05$) **(Figure 3.4)**. Initially the signal intensity of the feature K2 was plotted against the content of tanshinones. The graph obtained showed a

Table 3.2. Predicted locus/function of the 10 sequenced SDA features using blastN program through National Centre of Biotechnology Information (www.ncbi.nlm.nih.gov). Showing the best match as the putative identity for each sequence. E-value regarded as significant if < 1e-10. NA indicates the absence of significant data.

Feature ID	Length (bp)	Matching database entry	Putative identity	E Value
A5	526	-	No hits	NA
A11	492	-	No hits	NA
B7	218	-	No hits	NA
C11	492	-	No hits	NA
G9	341	-	No hits	NA
H17	612	-	No hits	NA
I5	398	AY757816.1	*Acorus gramineus psaI* gene, complete cds; and *ycf4* gene, partial cds; chloroplast	5e-45
		DQ673256.1	*Forsythia europaea psaA-psbB* fragment chloroplast	1e-148
K2	406	-	No hits	NA
L5*	407	AF193940.1	*Digitalis purpurea* 18S ribosomal RNA (nSSU) gene, partial sequence†	0.0
O14*	504	GQ997211.1	*Dillenia indica* voucher FLAS:M.J. Moore 340 putative RF1 protein (*ycf*1) gene, complete cds; chloroplast†	0.0

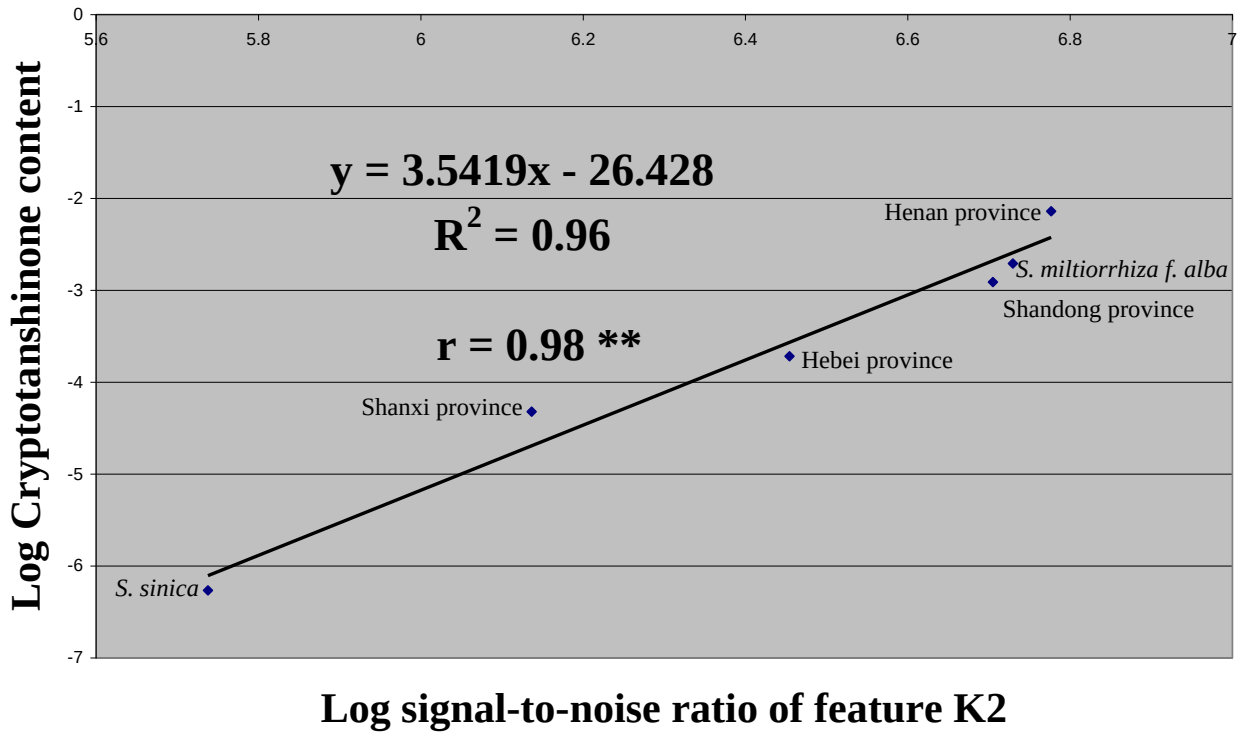
* Features that were found positive for the driver target.
† This feature also matched perfectly to other numerous species.

Table 3.3. Correlations among the signal of 8 most discriminatory features and the agronomical traits.

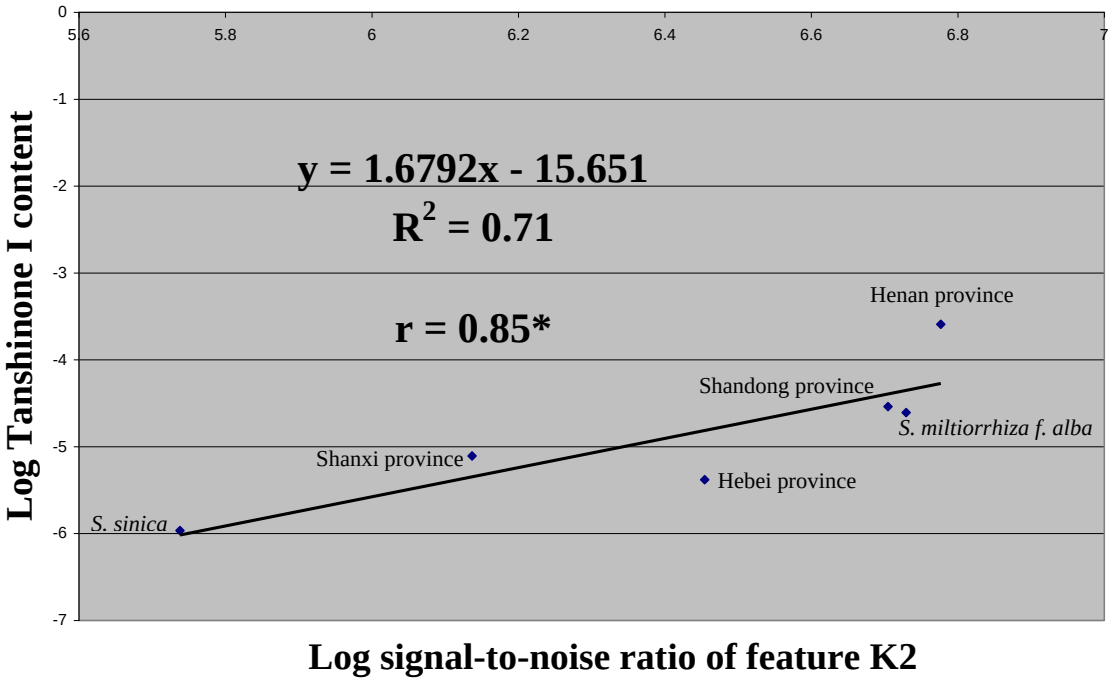
Signal / Agronomical Traits	C11	K2
Number of side branches per plant-pair		
Aerial fresh weight (g/plant-pair)	**-0.87*** **0.02**	
Root number per plant-pair	**-0.97**** **0.00**	**-0.86*** **0.03**
Maximum root diameter (mm)		**-0.91*** **0.01**
Root fresh weight (g/plant-pair)	**-0.86*** **0.02**	**-0.86*** **0.03**

** Correlation is significant at the 0.01 level
*Correlation is significant at the 0.05 level

A) **Correlation among the content of Cryptotanshinone and K2 signal strength.**



B) **Correlation among the content of Tanshinone I and K2 signal strength.**



C) **Correlation among the content of Tanshinone IIA and K2 signal strength.**

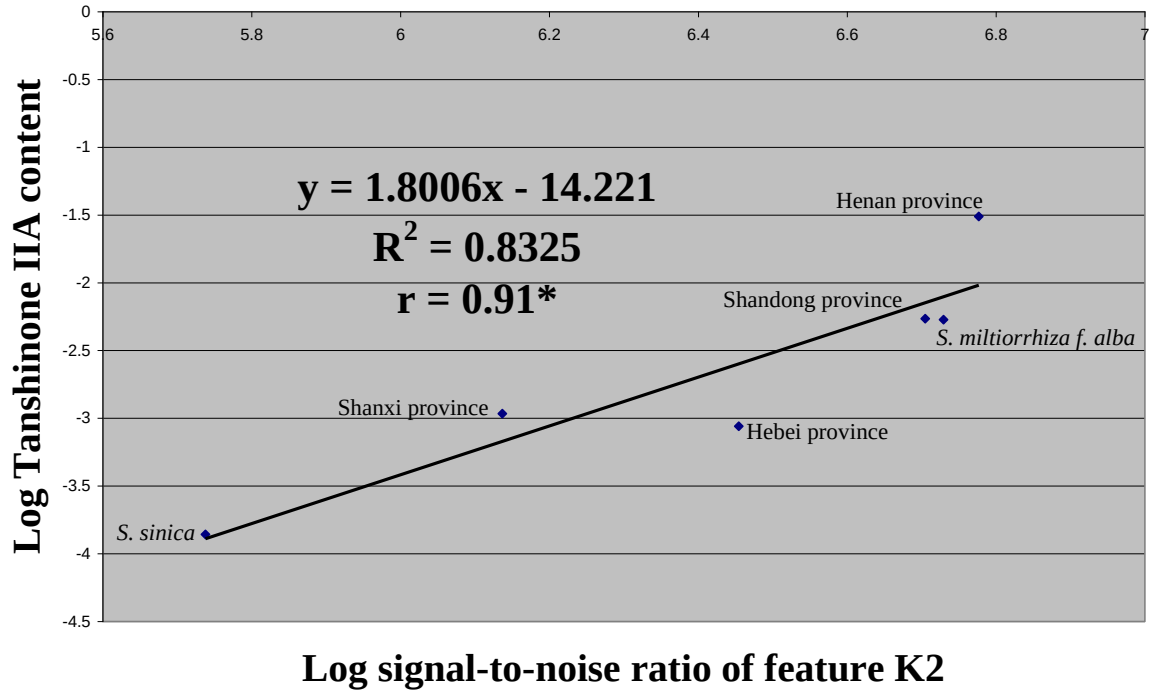


Figure 3.4. Significant correlations among the contents of three tanshinones and the signal of feature K2.
** Correlation is significant at the 0.01 level
* Correlation is significant at the 0.05 level

potential curve (data not shown) and therefore these data was transformed using the logarithm of base 2 in order to linearize it prior to performing the linear regression analysis. As it can be seen **Figure 3.4** shows a linear correlation where the hybridization signal of K2 and content of all three tanshinones were the lowest for *S. sinica* and the highest for the populations of the provinces of Henan, Shandong and *S. miltiorrhiza f. alba* (Shandong province). Consequently, it can be implied that the hybridization signal of K2 have the same pattern of variation as the content of the three tanshinones in the populations analyzed.

3.4 DISCUSSION

This chapter demonstrates that *Salvia*-SDA was able to fingerprint geographical populations of *S. miltiorrhiza*. The data obtained was useful for establishing genetic similarities among *S. sinica* and five populations of *S. miltiorrhiza*. Moreover, the subset of the most discriminatory features identified was used for correlation analysis with morphological and chemical profiles obtained by previous studies. In this following section the usefulness of the SDA, the fingerprinting procedure and genetic diversity determination within and among populations of *S. miltiorrhiza* is discussed. Additionally, the prospect of identifying possible potential markers that could assist in marker assisted selection of *S. miltiorrhiza* is explored.

3.4.1 Genetic variation within and among five populations of *S. miltiorrhiza* and one of *S. sinica*

S. sinica was clearly differentiated from the five populations of *S. miltiorrhiza* in both hierarchical cluster analyses **(Figure 3.1 and 3.3)**. This result corroborate the findings in Chapter 2 **(Section 2.4.3),** implying that this array could be useful for authentication purposes. However, further studies are needed in order to establish the usefulness of the SDA for the identification of adulterated samples. For example, an angiosperm-specific SDA has been successfully used for detecting deliberate adulteration of dried commercial samples of ginseng, where *Panax quinquefolius* was added to *Panax ginseng* at a ratio of 1:10 (Niu et al., 2011b).

The hierarchical clustering also shows that there is a close genetic similarity between the populations of *S. miltiorrhiza*. Other molecular studies were able to fingerprint *S. miltiorrhiza* populations using different molecular markers such as EST-SSR, RAPD, ISSR, SRAP and CoRAP. Despite the fact that almost all studies had employed a different set of populations; they agreed that there is a close genetic similarity among them (Deng et al., 2009; Guo et al., 2002; Song et al., 2010; Wang et al., 2009). However, the results from previous studies also indicate that the major portion of total variation of *S. miltiorrhiza* existed within populations and that minor variations were found among populations (Guo et al., 2002; Song et al., 2010). The analyses used in this study could not provide an estimation of the genetic variation within *S. miltiorrhiza* populations since both hierarchical clusters obtained with the 285 and 8 features **(Figures 3.1 and 3.3)** were based on representations made from a DNA pool of five

different plants and not from individual plants. Moreover, the use of pooled DNA samples to assess the genetic diversity among populations may possess some inherent disadvantages. For example, it has been found that genetic information may be lost during pooling since alleles present at low frequencies may not be detected (Chuang et al., 2010b). Therefore, the genetic variations within the geographical populations employed in this study should be determined in order to find if pooling could have biased the distances among the populations by reducing or homogenizing the within-accession DNA variations.

A subsequent study on genetic diversity within *S. miltiorrhiza* populations conducted in our laboratory was able to present an estimation of the genetic variation within two of the same populations studied (Lau et al., 2011). This study was able to obtain the fingerprints of each of the five plants that constitute the pool of Henan and Shanxi province together with the fingerprints of the pooled samples (*S. sinica* and the five populations of *S. miltiorrhiza*) by using the same *Salvia* specific SDA. The data obtained was used to construct a dissimilarity dendrogram also based on the 285 features **(Appendix 11)**. The dissimilarity dendrogram produced four clusters at the cut off-point of 5. As may be seen in **Appendix 11**, cluster 1 contained the *S. miltiorrhiza* samples that were constructed by pooling of the individual plants with the exception of *S. miltiorrhiza f. alba* from Shandong province which was in cluster 2. Cluster 1 also contained most of the individual plants assessed, however plants 4 and 5 from Shanxi province and plant 4 from Henan province clustered separately (Cluster 3). The clear differentiation between plants 4 and 5 from Shanxi when compared to other plants from same province may indicate a high level of variation within the population from this

province. The same differentiation was also found within the individuals of Henan province. Therefore, it may be possible that pooling of the five plants had reduced or homogenized the DNA variations within Shanxi and Henan provinces, which at the same time could have biased the distances given among the populations. Consequently, the hierarchical clustering obtained with the pooled samples **(Figure 3.1)** could only give an estimation of the genetic distances between these populations.

The high level of genetic diversity found within populations may be mainly attributed to the fact that *S. miltiorrhiza* is predominantly cross-pollinated (Jin, 2001; Song et al., 2009b), this may explain an increase in genetic diversity since the level of heterozygosity in the populations would be increased. The samples of *S. miltiorrhiza* used in this study were pooled with the aim of obtaining a molecular profile for the populations that could be correlated to the agronomical and chemical data from previous studies (Li et al., 2009a) which was also obtained from pooled samples. However, due to the high level of genetic diversity within *S. miltiorrhiza*, pooling of different plants from the same population for chemical or molecular fingerprinting could produce misleading results since each plant could have a different profile that would not necessarily correlate with the pooled profile. Therefore, future studies on association between chemical and genetic variation should be performed on individual plants where each plant should be used for both analyses.

3.4.2 Possible cause of the reduced subtraction efficiency

The putative sequences of the two features that were found positive for the driver target **(Table 3.2)** helped to clarify the reasons for the reduced subtraction efficiency (**Section**

2.4.1). The sequence of feature L5 matched perfectly to a partial region of 18S rRNA gene of numerous species and the sequence of feature O14 matched perfectly to the putative RF1 protein (*ycf1*) gene sequence of numerous species. The results imply that these sequences are highly conserved in different plants, which explains why these features had strong hybridization signals for the driver target. However, these two features were found to be highly polymorphic across the populations of *S. miltiorrhiza*, which implies that the polymorphism present could be attributed to a difference in the copy number of these fragments. For example, changes in the copy number of the 18S-25S gene had been found not only in members of the same population of a single species but also among somatic cells of individual plants (Rogers and Bendich, 1987).

Therefore, these two sequences should have been eliminated during subtraction since they are highly homologous to the ones found in the driver. Although it is not possible to determine the exact cause of the inefficient elimination of these sequences, it may be possible to determine the steps in the suppression subtraction hybridization that were deficient. For instance, the sequences of these two fragments were found to possess different adaptors at each end, which may imply that ligation was properly performed. In addition, it implies that hybridization between homologous DNA strands possessing adaptor1 and adaptor 2R during the second hybridization was successful. The only likely explanation for the inclusion of these fragments is that during the first hybridization several of these adaptor-ligated sequences (with high copy numbers) remained single stranded instead of hybridizing with homologous sequences. This deficiency may be attributed to the fact that the quantity of driver added in the hybridization may have not been enough to hybridize with all the homologous

sequences present. Therefore, the adaptor-ligated sequences of these two fragments (L5 and O14) that remained single stranded may have hybridized with homologous adaptor-ligated sequences from the second tester pool during second hybridization. As a result, these fragments would have possessed different adaptors at their ends and could have been PCR amplified. In conclusion, it can be said that the presence of these fragments in the array may be evidence that one of the main reasons why the subtraction was not fully efficient was the lack of excess driver able to remove highly homologous sequences.

3.4.3 Identifying potential genotype specific markers among the most discriminatory features

The most discriminatory features could be potential markers for the genotyping of *S. miltiorrhiza* since they were found to be able to discriminate between the populations studied. Among the 8 features only I5 matched with a homologous sequence in GenBank **(Table 3.2)**. The putative sequence of feature I5 (*psaI* and *ycf4* gene) also had a significant alignment with *psaA-psbB* chloroplast region (DQ673256.1), which was also a match database entry for DNA sequences of features P4 and E13, that were part of the ten most discriminatory features in Chapter 2 **(Table 2.4)**. Although all three features matched the same database entry, they aligned to different segments of this *psaA-psbB* region which is 32,412 bp long (Lee et al., 2007). The previous study by Lee et al. (2007) found that this region have a series of inversions among members of the Oleaceae family (*Jasminum* and *Menodora*). Therefore, present and previous studies have shown that several segments of this *psaA-psbB* region are highly polymorphic

among different members of the Lamiales, which could make this region attractive as a potential genotyping marker for the Lamiales species.

The other seven unknown sequences **(Table 3.2)** may be potential markers that could assist in genotyping of populations. For instance, specific primers could be developed for amplification of these sequences within and among populations in order to find variations in amplification length and if possible to isolate and sequence these products in order to find which microstructural change is responsible for the observed polymorphism (insertion, deletion, tandem repeat, SSR, STR, inversion). Therefore these sequences may lead to the development of PCR-based markers for the genotyping of *S. miltiorrhiza*. Furthermore, some of these features could be potential molecular markers linked to desirable agricultural traits that could be useful for marker assisted selection of *S. miltiorrhiza*, as it is explained in the following section.

3.4.4 Correlations between the genetic and morphological/chemical profiles

To the best of this author's knowledge this is the first study to produce significant correlations between molecular profiles with agronomical traits and the content of tanshinones in *S. miltiorrhiza.* Other studies have concentrated on the relationship between the phylogenetic trees constructed based on ITS sequences and HPLC profiles of important bioactive components (including tanshinones), however no significant relationship has been found (Han et al., 2010; Xu et al., 2009).

There could be two important reasons why a significant correlation was found in this study and not in previous studies performed by Han et al. (2010) and Xu et al. (2009). Firstly, the previous studies constructed the phylogenetic trees based only on ITS sequences, which are part of nuclear ribosomal DNA. Nuclear ribosomal genes consist of individual repeats (18S-5.8S-26S) of which plant genomes have hundreds to thousands (Alvarez and Wendel, 2003). Therefore, there is a high possibility that any of these repeats could be linked to a gene responsible for the production of these bioactive compounds. However, other copies present in the genome could be found in other chromosomal loci, having no linkage with the gene, thus decreasing the correlation coefficient of this sequence with the chemical content. Therefore, the high copy number of this marker system may be a disadvantage for association studies for a desirable trait. Secondly, the correlation analyses of the two previous studies were performed with methods that may not have been the most appropriate. For instance, Han et al. (2010) found a relationship between the molecular profiles obtained by ITS2 sequences and HPLC profiles of different *Salvia* species; however it was not statistically significant. The correlation was performed with partial least squares (PLS) which has been found in another study which employed this method for correlating AFLP and HPLC profiles in *Echinacea purpurea* to be inappropriate (Baum et al., 2001). Instead, Baum et al. (2001) was able to find correlations using individual regressions, Mantel's test and canonical correlation analysis. However PLS did not produce a set of components among DNA variables that yielded good linear models for all chemical variables.

In the present study, the correlation analysis was performed with Pearson bivariate correlation and individual regressions which have been successfully employed to

correlate molecular and chemical profiles in other studies (Baum et al., 2001; Chen et al., 2009a). The correlation analysis found that signal intensity of feature K2 was negatively correlated to root morphology and positively correlated to tanshinone content. These findings are in agreement with Li et al. (2009a), who found that root diameter and tanshinone content are inversely correlated, implying that plants with small roots have higher content of tanshinones. The results also imply that the hybridization signal of K2 have the same pattern of variation as the content of the three tanshinones, since the higher the hybridization signals of K2 the higher the content of tanshinones at least in the populations analyzed. Consequently, the K2 sequence may be a potential marker for tanshinone content. However, since this feature did not match with any known DNA sequence or protein translations in GenBank, it is unknown which polymorphisms were associated with the variation in the K2 signal intensity among the populations. Therefore, further analyses have to be performed before it is possible to confirm the use of K2 as a marker. Firstly, it will be necessary to determine the association of K2 signal with tanshinone content using other populations of *S. miltiorrhiza*. It is important to note that the plants used for HPLC and morphological analyses were different to the ones used for the hybridization analyses. They were sourced from the same pool of seeds collected from the same populations; however different plants were used for each study. Also, the results presented were obtained by pooling different plants from the same population. Therefore, in order to determine if K2 is a good predictor of tanshinone content, chemical and molecular analysis should be performed on individual plants and the same plant should be used for both analyses. Secondly, it will be important to find which K2 polymorphisms are associated with high or low tanshinone content. For instance, specific primers could be designed to amplify

the K2 locus within and among populations in order to find variations in amplification length and if possible to isolate and sequence these products. The sequences obtained could be aligned to reveal which microstructural changes are present in this locus. Furthermore, the full length of the genome of *S. miltiorrhiza* has been already sequenced by the Institute of Medicinal Plant Development (IMPLAD) using next-generation sequencing (Chen, 2010). Although, the complete sequences are not available on public databases, future studies could locate the K2 locus in the genome to find if it is linked to coding regions that are involved in tanshinone production.

If the K2 locus is found to be a good predictor of tanshinone content, then the derived equations from the correlations could be used to predict the contents of tanshinones in seedlings. Similar studies performed in *Echinacea purpurea* have shown that regression equations obtained from the correlation of RAPD markers with phytochemical traits provided a good prediction of total phenol content in aerial parts of the plant (Chen et al., 2009a). Therefore, if further efforts are made to standardize environmental and agricultural conditions in which the plants are grown, K2 has potential to be used in plant breeding programs as the plants may be screened at the seedling stage. This is an important advantage when compared to HPLC analysis where mature roots are used to assess tanshinone content.

3.5 CONCLUSIONS

SDA has demonstrated to be an efficient technique for fingerprinting S. *sinica* and geographical populations of *S. miltiorrhiza*. Previous studies have obtained similar results using ITS sequence comparisons as well as chemical profiles (Han et al., 2010; Xu et al., 2009). However, to the best of author's knowledge, no other study has found a significant correlation between chemical and molecular profiles in *S. miltiorrhiza*.

Although SDA has been shown to be an efficient technique for fingerprinting *S. miltiorrhiza* populations, the hierarchical subdivision obtained with the pooled samples could only give an estimation of the genetic diversity between these populations since the pooling could have affected the outcome of the analysis by homogenizing the within-population DNA variations. Therefore, future studies on association between chemical and genetic variation should be performed on individual plants and the same plant should be used for both analyses since pooling of different plants from the same population for chemical or molecular fingerprinting could produce misleading interpretations.

Furthermore, the SDA was able to identify DNA sequences that are highly polymorphic among *S. miltiorrhiza* geographical populations. Most of these highly polymorphic sequences were unknown sequences that could be potential markers used for genotyping of *S. miltiorrhiza*. Finally, if future studies are able to confirm the significant correlation found between the K2 feature with root morphology and the content of tanshinones in *S. miltiorrhiza*, K2 locus could be used of as a potential marker for predicting tanshinone content.

CHAPTER 4

Fingerprinting of *Echinacea* species using the *Echinacea* Subtracted Diversity Array

4.1 INTRODUCTION

This chapter describes the use of suppression subtractive hybridization to enrich selectively the SDA with polymorphic and divergent DNA sequences from the genus *Echinacea*. Additionally, describes how this *Echinacea*-specific array was used for fingerprinting different species and accessions of this genus. Finally, the SDA genetic profiles were correlated with chemical profiles obtained from previous studies.

Echinacea (Asteraceae) is a North American genus which is widely known for their medicinal uses **(Section 1.2.2)**. The number of taxa of this genus has differed significantly (from 2 to 11 groups) depending on the nomenclature (Baum et al., 2004). The taxonomical classification that is most commonly used is that of McGregor, established in 1968, which recognized nine species and four varieties **(Table 1.2)** (McGregor, 1968). A more recent taxonomic revision based on morphometric analyses (Binns et al., 2002a) recognized four species and eight varieties **(Table 1.2)**. Although, results from other investigations do not completely support either of these classifications (Kim et al., 2004; Mechanda et al., 2004a; Wu et al., 2009), McGregor's classification is more widely used botanically and commercially. This classification will also be employed in this chapter.

Molecular fingerprinting studies were performed with the aim to elucidate the genetic relationships for *Echinacea* and to find independent support for the morphology-based classifications. For instance, AFLP markers were employed to fingerprint all species in the genus (Kim et al., 2004; Mechanda et al., 2004a). The study performed by Kim et al. (2004) found two major clades, one containing *E. purpurea* (L.) Moench., *E. sanguinea* Nutt. and *E. simulata* McGregor. and the other containing the remaining species. Their data indicated that all *Echinacea* taxa are closely related sensu McGregor. In contrast, Mechanda et al. (2004a) found support for the four species classification of Binns et al. (2002a) but not for the varieties. The discrepancy between these two studies could be attributed mainly to the primer combination used, the number of individuals sampled and to the fact that Kim et al. (2004) used a neighbor-joining algorithm to construct the tree. Phylogenetic studies have also been performed for this genus; however they were unable to resolve the species-level relationships due to the low levels of molecular divergence found in the selected loci. For instance, the sequence divergence of two chloroplast (*trn*S and *trn*G) and three nuclear loci (*Adh* (alcohol dehydrogenase), *Ces*A (cellulose synthase) and *GPAT* (3-phosphate acetyltransferase) were unable to provide a resolved topology or congruent hypotheses about species-level relationships (Flagel et al., 2008). In addition, sequence divergence of ITS sequences and intervening 5.8S regions found that the sequence divergence within *Echinacea* species ranged from 0.18% to 3.2% and several species had identical ITS-2 sequences (Urbatsch et al., 2000). Consequently, the results of the different molecular studies are also contradictory which makes it difficult to resolve completely the relationships among the species.

In order to clarify the genetic relationships within this genus, there is a need for molecular techniques that are not only able to distinguish species and varieties but which are also able to overcome the main limitations of PCR-based techniques, i.e., the assumption that comigrating fragments are homologous. A previous study using AFLP have found that comigrating polymorphic bands from different species and varieties of *Echinacea* were not homologous (Mechanda et al., 2004b). The sequence identity of the polymorphic fragments ranged from 23 - 64% within variety, from 23-46% within species and as low as 1.25% within the genus. This is a significant disadvantage since the data obtained from these techniques are usually inappropriate for phylogenetic studies. SDA could potentially be a superior technique for assessing the genetic relationships among *Echinacea* species since it does not require previous DNA sequence information and has shown to be capable of differentiating between closely-related species and populations such as the ones for *S. miltiorrhiza* (Chapter 2 and 3). Although the SDA has not been used previously for phylogenetic purposes, it could be used to clarify relatedness of *Echinacea* species. Furthermore, the molecular profile obtained with the SDA could be employed for the identification of potential molecular markers that could be genotype specific or that could be associated with bioactive compound content. There is a recent study performed by (Wu et al., 2009), in which metabolites of roots were analyzed by HPLC-photo diode array (HPLC-PDA), GC-MS, and multivariate statistical methods. The lines analyzed represented a broad geographical and morphological sampling and were also used in the phylogenetic study described above (Flagel et al., 2008). It will be of interest to use these same lines to develop the molecular profiles with the aim of identifying potential molecular markers

associated with the production of bioactive compounds and to compare if there is any resemblance among the dendrograms obtained with molecular and chemical analyses.

The objectives of the experiments described in this chapter were: (1) to generate a SDA enriched for polymorphic and divergent DNA sequences for *Echinacea,* (2) to evaluate the potential of the SDA to fingerprint *Echinacea* species and to assess the genetic relationships among them, (3) to establish if the differences found in the content of bioactive compounds previously studied could be related to the genetic profiles, and (4) to identify potential molecular markers useful for species authentication.

4.2 MATERIALS AND METHODS

4.2.1 Plant material

In order to develop a gDNA representation for the subtraction, the DNA from a total of 142 species including angiosperms and non-angiosperms were soured as described in **Section 2.2.1.** In addition, a total of 24 lines were used to represent the *Echinacea* genus. Five *Echinacea* species, sensu McGregor, (*E. angustifolia* DC., *E. paradoxa* (Norton) Britton, *E. pallida* (Nutt.) Nutt., *E. purpurea* and *E. tennesseensis* (Beadle) Small.) were obtained from three different sources; the other four species (*E. atrorubens* Nutt., *E. laevigata* (Boynton & Beadle) Blake, *E. sanguinea*, *E. simulata*) could not be obtained as quarantine restrictions applied to these species**.** Nineteen of the 24 lines were selected from the germplasm collection of the U.S. National Plant Germplasm System maintained by the USDA-ARS North Central Regional Plant Introduction

Station (NCRPIS) **(Table 4.1)**. These 19 lines had been previously used in two previous studies by Flagel et al. (2008) and Wu et al. (2009). The other remaining lines were obtained from Botanical Resources Australia (Tasmania) and from verified specimens from a specialized plant nursery (The Diggers Club, Dromana Victoria) **(Table 4.1)**.

4.2.2 Construction of the *Echinacea* Subtracted Diversity Array

4.2.2.1 DNA extraction and development of tester and driver pools

Genomic DNA was extracted from fresh leaves using a modification of the standard CTAB procedure (Doyle and Doyle, 1987) and subsequently clean up using the DNeasy® column of the DNeasy® Plant Mini Kit (Qiagen) (**Section 2.2.2.1**). The genomic DNA from *Echinacea* lines sourced from the U.S. National Plant Germplasm System and Tasmania were extracted from seedlings (approximately 0.3 g) using the same protocol.

All DNA samples were pooled as described in **Section 2.2.2.1** to obtain representations of the following seven groups: *Echinacea* (subtraction pool), Asterids (excluding Asteraceae) **(Table 4.1)**, non-angiosperms, Monocots, Magnoliids, Rosids, and Eudicots not belonging to the Rosids or Asterids (Eudicots and Core Eudicots) **(Table 2.1)**. The plants used to represent the Asterid clade were different from those used in the *Salvia* study. In this study species belonging to the Lamiaceae were included, whereas those closely related to *Echinacea* (belonging to the Asteraceae) were excluded.

Table 4.1. Description of the Asterids and *Echinacea* species used for DNA extraction and development of genome representations.

REPRESENTATIONS	SPECIES	
ASTERIDS (36 species)	*Angelica archangelica* *Angelica dahurica* *Bacopa monnieri* *Camellia sinensis* *Centella asiatica* *Coffea arabica* *Digitalis purpurea* *Forsythia suspensa* *Glechoma hederacea* *Hedeoma pulegioides* *Hyssopus officinalis* *Ilex paraguariensis* *Impatiens* sp. *Leonourus sibiricus* *Leonurus cardiaca* *Lycium barbarum* *Melissa officinalis* *Mentha* × *piperita* *Mentha arvensis var. piperascens*	*Malinv. ex L. H. Bailey* *Mentha pulegium* *Mentha spicata* *Nepeta cataria* *Perilla frutescens* *Plantago major* *Platycodon grandiflorus* 'Apoyama' *Prunella vulgaris* *Sambucus nigra* *Scrophularia nodosa* *Scutellaria lateriflora* *Stachys officinalis* *Symphytum officinale* *Thymus vulgaris* *Valeriana officinalis* *Verbascum thapsus* *Vitex agnus-castus* *Withania somnifera*
***Echinacea* (5 species) Taxon (sensu McGregor)**	**SUBTRACTION POOL**	**(total of 24 accessions)[ad]**
	***E. angustifolia* DC.** PI631267 (OK).ang 267	***E. pallida* (Nutt.) Nutt.** PI631275 (OK).pal 275 PI631290 (IA).pal 290 PI631293 (AR).pal 293 PI631296 (MO).pal 296 PI631315 (NC).pal 315
	E. angustifolia* DC. var. *angustifolia PI631272 (OK).ang-ang 272 PI631285 (IA).ang-ang 285 PI631318 (KS).ang-ang 318	***E. purpurea* (L.) Moench** PI631307 (MO).pur 307 PI631313 (NC).pur 313 PI633669 (LA).pur 669
	***E. angustifolia* DC. var. *strigosa* McGregor** PI631266 (OK).ang-str 266 PI631320 (OK).ang-str 320	***E. purpurea*** "Double Decker"[b] "White purpurea"[b] "purpurea"[b]
	E. paradoxa* (Norton) Britton var. *paradoxa PI631301 (MO).px-px 301 PI631321 (MO).px-px 321	***E. pallida*** "Hula dancer"[b].
	***E. paradoxa* var. *neglecta* McGregor** PI631263 (OK).px-neg 263 PI631264 (OK).px-neg 264 PI631265 (OK).px-neg 265	***E. tennesseensis*** [c]

	NOT INCLUDED IN THE SDA DEVELOPMENT
Echinacea	**Putative hybrid** ***E. paradoxa* var. *paradoxa* and *E. pallida*** PI631294 (AR).hyb 294
	E. angustifolia Plot 9[c] (OR).ang plot 9
	E. pallida Plot 5[c] (Germany).pal plot 5
	E. purpurea Plot 10009[c].pur plot 10009 (Commercial crop)

Notes: AR, Arkansas; IA, Iowa; KS, Kansas; LA, Louisiana; MO, Missouri; NC, North Carolina; OK, Oklahoma; OR, Oregon; SC, South Carolina; TN, Tennessee; VA, Virginia.

[a] *Echinacea* with PI accessions numbers were obtained from the germplasm collection in the U.S. National Plant Germplasm System.
[b] *Echinacea* verified specimens obtained from a specilaized plant nursery (The Diggers Club. Dromana VIC).
[c] *Echinacea* obtained from the Botanical Resources Australia (Tasmania).
[d] The abbreviated names in blue are the names used to refer to the lines in the figures and tables.

4.2.2.2 Subtraction

Subtraction was performed using the PCR-Select™ Bacterial Genome Subtraction Kit (Clontech), following the manufacturer's protocol. The protocol was slightly modified as described below:

Firstly, equal amounts of DNA extracted from the 24 *Echinacea* lines were bulked to form the tester pool. The driver pool was formed by bulking 700 ng of each representation with the exception of the *Echinacea* pool **(Table 4.1)**. Digestion and purification of driver and tester pool, were performed as described in **Section 2.2.2.2.**

Secondly, during adaptors ligation (Adaptors 1 and 2R) to the digested *Echinacea* pool, 1.6 ng of human skeletal muscle cDNA (control) was added to each of the reactions before ligation. This was deliberately performed in order to monitor the efficiency of the ligation. Human skeletal muscle cDNA was obtained from the PCR-Select™ cDNA Subtraction Kit (Clontech) and ligation was verified as described in **Section 2.2.2.2. Figure 4.1** shows a product of about 750 bp for samples 1 and 3, which confirms the ligation was successfully performed. The human skeletal muscle cDNA was subsequently removed from the *Echinacea*-specific DNA during hybridizations by adding a total of 26 ng of this cDNA in the driver.

Thirdly, the two hybridizations were performed at 68^0C employing a tester:driver ratio of 1:60. A higher excess of driver was employed compared to the one used for the *Salvia* subtraction (1:30), to ensure the subtraction of all the homologous sequences

between the tester and driver as these homologous sequences were not entirely eliminated during the *Salvia* subtraction **Section 3.4.2**.

Finally, primary and secondary PCR were performed on the product obtained after second hybridization in order to exponentially amplify *Echinacea*-specific DNA as described in **Section 2.2.2.2**. Unexpectedly, the patterns of the secondary PCR product for subtracted and unsubtracted were almost the same **(Figure 4.2)**. However the expected bands were clearly seen in the control, indicating that the subtraction was successful.

4.2.2.3 Cloning of the subtracted sequences

Amplification, purification and cloning of the *Echinacea*-specific DNA was performed as described in **Section 2.2.2.3**.

4.2.2.4 Microarray construction and printing

Microarray construction and printing was performed as described in **Section 2.2.2.4.** The protocol was slightly modified as described below:

The template used for the amplification of the cloned inserts was not the purified plasmid obtained after using the Miniprep Kit (Qiagen). Instead, 10µl of bacterial cell culture was mixed with 10 µl of MilliQ water and then heated at 100^0C for 10 min to disrupt the cells and release the plasmid DNA. Then 1.5 µl of this sample was used as template to amplified the cloned inserts using nested primers 1 and 2R (Clontech).

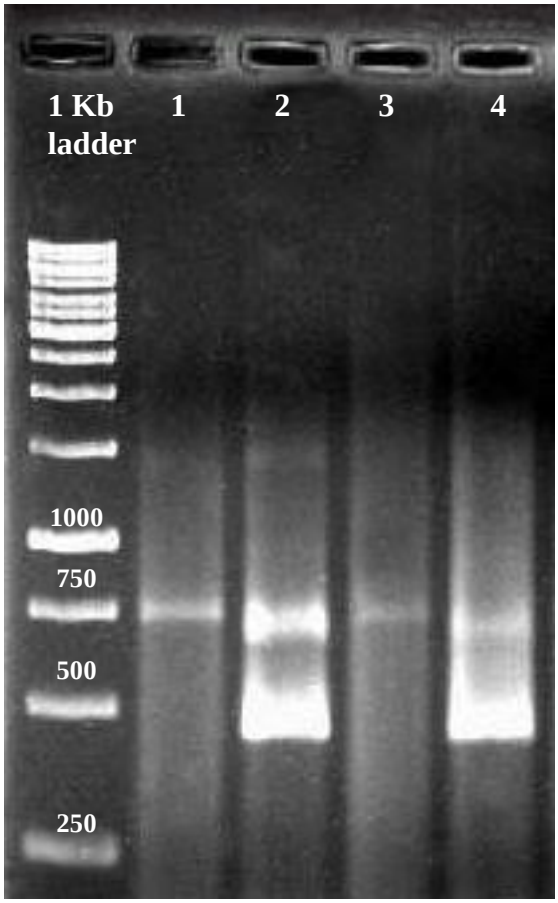


Figure 4.1 Results of the ligation efficiency analysis

Lane 1. PCR products using digested *Echinacea* pool ligated to **adaptor 1** with added cDNA skeletal muscle as a template. **G3PDH 3' primer and PCR Primer 1**.

Lane 2. PCR products using digested *Echinacea* pool ligated to **adaptor 1** with added cDNA skeletal muscle as a template. **G3PDH 3' and 5' Primers.**

Lane 3. PCR products using digested *Echinacea* pool ligated to **adaptor 2R** with added cDNA skeletal muscle as a template. **G3PDH 3' primer and PCR Primer 1**.

Lane 4. PCR products using digested *Echinacea* pool ligated to **adaptor 2R** with added cDNA skeletal muscle as a template. **G3PDH 3' and 5' Primers.**

1.5% agarose/EtBr gel.

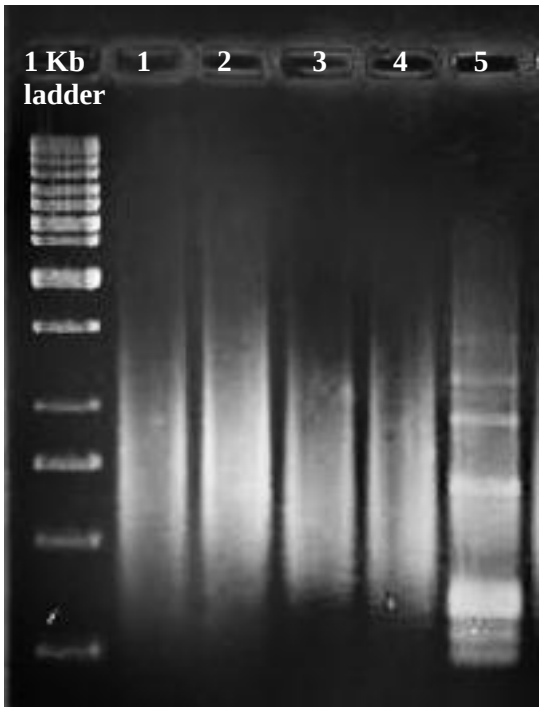


Figure 4.2. Secondary PCR products of the subtracted *Echinacea* pool

Lane 1 and 2. Subtracted *Echinacea pool*.
Lane 3 and 4. Unsubtracted *Echinacea* pool.
Lane 5. PCR control subtracted cDNA provided with the kit.

1.5% agarose/EtBr gel.

A total of 283 clones together with 17 controls **(listed in Appendix 12)** were amplified, precipitated and printed. The parameters used to print are found in **Appendix 2**. Ten subarrays (each subarray composed of 283 clones and 17 controls) were gridded on aminosilane-coated slides.

4.2.3 Validation of the array and fingerprinting of the *Echinacea* lines

The SDA was firstly validated by performing separate hybridizations with the biotin-labeled DNA from the *Echinacea* and driver pools. Secondly, fingerprints were obtained by hybridizing biotin-labeled DNA of each *Echinacea* line to the array. Labeling of the targets and hybridizations were mainly performed as described in **Sections 2.2.3.1 and 2.2.3.2**. However, slight modifications were performed:

Firstly, in order to compare the subtraction efficiency between this array and the previous *Salvia* array, hybridizations for the *Echinacea* and driver pools were performed at 42^0C.

Secondly, during the fingerprinting of the *Echinacea* lines, hybridization of the biotin-labeled DNA of each *Echinacea* accession was performed at 47^0C, instead of the 42^0C used for all *Salvia* experiments, to facilitate a higher level of discrimination between the lines. In addition, it should be noted that all hybridizations were performed with five technical replicates (subarrays) and two biological replicates, for a total of ten data points per array feature.

Thirdly, the temperatures and times of the four stringency washes were also modified to obtain higher level of discrimination between the lines: The slides were washed once in 1 x SSC, 0.1% SDS at 37^0C for 8 min, once 1 x SSC, 0.1% SDS at 40^0C for 5 min, once in 0.1 x SSC, 0.1% SDS at 35^0C for 5 min and once in 0.1 x SSC at 35^0C for 5 min.

Finally, a total of 27 lines were fingerprinted; 23 of which were used to construct the subtraction pool (excluding *E. purpurea* "Double Decker") and four additional lines which were not employed in the SDA construction **(Table 4.1)**. **)**. The reproducibility found for the biological replicates of these fingerprints was higher than 0.97, after optimization of the hybridization conditions.

4.2.4 Analysis of the *Echinacea*-array

4.2.4.1 Scanning, quantification and data analysis

Scanning of the slides, quantification and subsequent analysis (background correction, flagging, mean of the signal intensity across the five technical replicates, normalization and combination of the biological replicates) were performed as described for the fingerprinting and data analysis of *Salvia* species **(Section 2.2.5)**. The normalized mean signal intensity was then used for all subsequent analyses; however the efficiency of subtraction was calculated using binomial scoring in order to compare the efficiency of the subtraction with previous studies. The cut-off point chosen was based on the hybridization signal of a specific control feature present in the array. This control was an aliquot of the enriched *Echinacea*-specific sequences obtained from the subtraction

process prior to cloning **(Appendix 12)**. Any signal intensity lower or equal to the signal intensity of this control for the driver target was considered a negative spot.

4.2.4.2 Statistical analysis

Histograms for the raw signal intensities of tester and driver pools were constructed as described in **Section 2.2.5.3.** In addition, PCA, the magnitude of variance and Pearson bivariate correlation were used to identify a subset of the most discriminatory features with unique patterns of variation among the 283 features in the array. Finally the normalized mean signal of the full feature set and a subset of the most discriminatory features were used to perform separate hierarchical cluster analyses (**Section 2.2.5.3**).

Data obtained from a previous study (Wu et al., 2009) on metabolomic profiling of *Echinacea* genotypes was used for correlation analyses with the hybridization data. The relative abundance of 43 lipophilic metabolites in roots from 6-month-old plants was correlated with the normalized mean signal of the full feature set by performing Pearson bivariate correlations (SPSS version 17.0) and regression analysis (Microsoft Excel). The correlations were performed for only **19 lines** that were shared by the two studies.

4.2.5 Sequencing of selected polymorphic features

Amplification products of the highly polymorphic features were sequenced and nucleic acid and protein homology searches were performed using blastN and blastX programs through the National Center of Biotechnology Information (www.ncbi.nlm.nih.gov) as described in **Section 2.2.6.**

4.3. RESULTS

4.3.1 Subtraction efficiency and validation of the microarray

Eight (3%) positive features were found after hybridizing the driver target with the array, indicating that the subtraction procedure was able to isolate *Echinacea*-specific DNA sequences with 97% efficiency. A histogram was also constructed with the signal intensities obtained after hybridizations of the tester and driver pools onto the SDA. As indicated in **Figure 4.3**, the distribution of the signal intensity of the tester overlaps with that of the driver between the signal range of 0 and 10. Comparing these results with the histogram obtained for *Salvia* **(Figure 2.7)**, where the distribution of the tester overlaps the signal of the driver between 0 and 60, it may be implied that the *Echinacea* array had fewer features that gave high signals for the driver pool when compared to the previous *Salvia* array. Based on the above results, it may be implied that the *Echinacea* array had a lower percentage (3%) of sequences homologous to the driver which may represent the non-subtracted sequences.

.

4.3.2 Fingerprinting of twenty-seven *Echinacea* lines

Fingerprints for twenty-seven *Echinacea* lines were obtained, which were representative of five species (sensu McGregor) (*E. angustifolia*, *E. paradoxa*, *E. pallida*, *E. purpurea* and *E. tennesseensis*). Out of the twenty-seven, four fingerprints corresponded to lines (Plot 9, 5, 10009 and accession PI631294) which were not used in the construction of the original subtraction pool from which the subtraction technique was performed **(Table 4.1)**. Representative photographs of the fingerprints can be seen in Appendix 13.

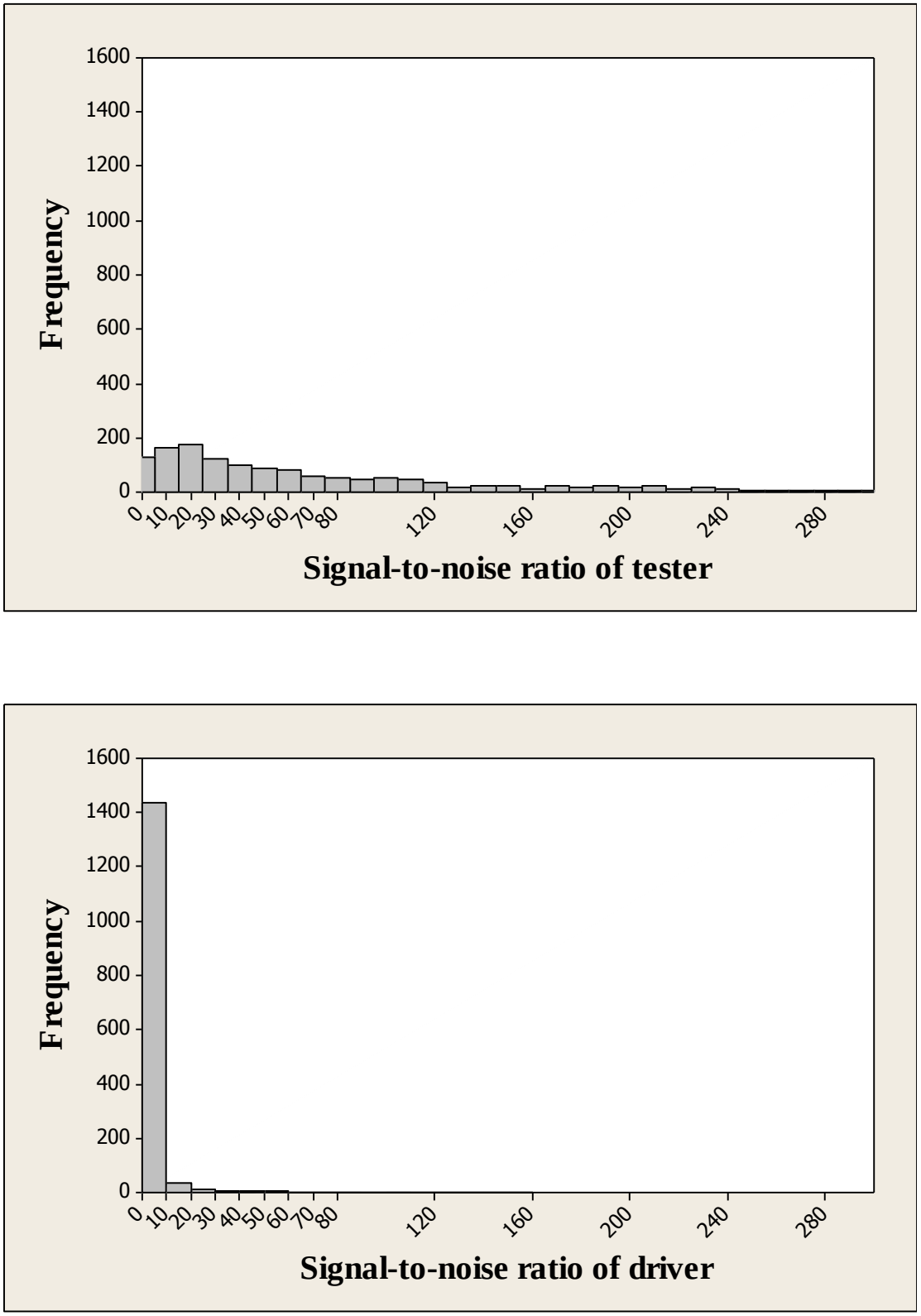


Figure 4.3. Histogram of the signal intensities obtained after hybridizations of the tester and driver pools.

The hierarchical cluster analysis constructed with the signal intensities of 283 features provided a clear differentiation between the twenty-seven *Echinacea* lines **(Figure 4.4)**. This dissimilarity dendrogram produced ten clusters at the cut off-point of 5. Cluster 1 included all lines belonging to *E. paradoxa*. Cluster 2 contained two lines of *E. pallida*, two lines of *E. angustifolia* and *E. tennesseensis.* Cluster 3 and 4 included lines from *E. angustifolia* and *E. pallida*. Cluster 5 consisted of only the putative hybrid of *E. paradoxa* var. *paradoxa* and *E. pallida*. Clusters 6 through 10 contained all lines belonging to *E. purpurea*.

Principal component analysis indicated that a high percentage of the variation (96.9%) may be explained by the first two components. The first principal component accounted for 94.7% of the variation and the second component explained only 2.2% of the variation **(Figure 4.5)**. In addition, it was observed that features that clustered close to zero had low variances among the populations while the features that were distributed throughout the plot were features that presented the highest variances **(Table 4.2)**. Based on this analysis, only the six most distant features from zero on the X axis were chosen since the first component explains most of the variation. Finally, the loading plots obtained also from the PCA analysis did not agree with the hierarchical cluster analysis. The plots did not show any clear differentiation across the species. Only three *E. purpurea* lines formed a separate cluster and four *E. paradoxa* lines cluster closer to the zero on the X axis **(Appendix 5)**.

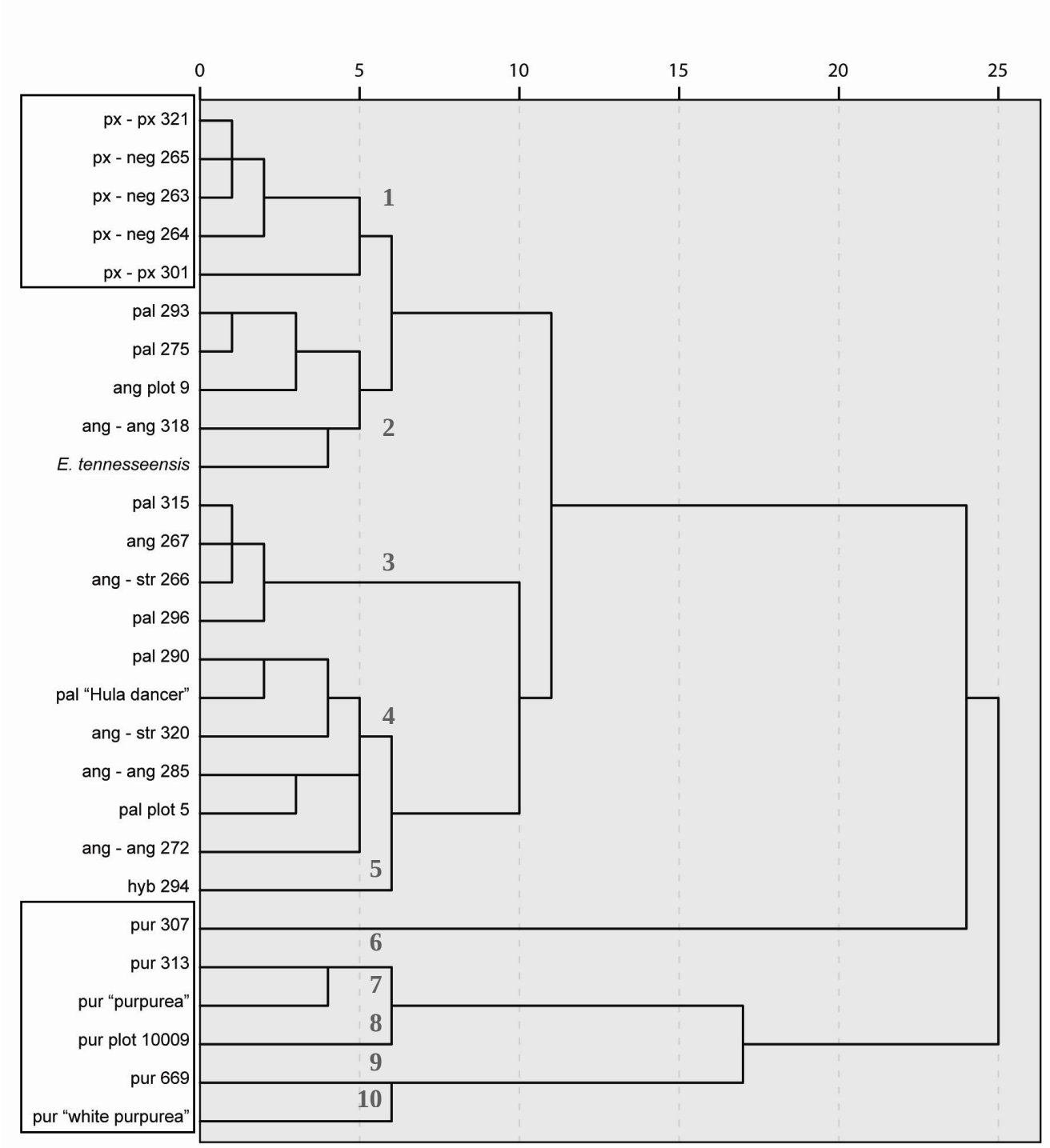


Figure 4.4. Dissimilarity dendrogram (Squared Euclidian distance, between groups linkage) for the SDA hybridization patterns of the 27 genotypes using the 283 features. The steps of the dendrogram show the combined clusters and the values of the distance coefficients at each step; the values have been rescaled to numbers between 0 and 25, preserving the ratio of the distances between the steps. The equivalents of the abbreviated names used for each of the lines are shown in Table 4.1.

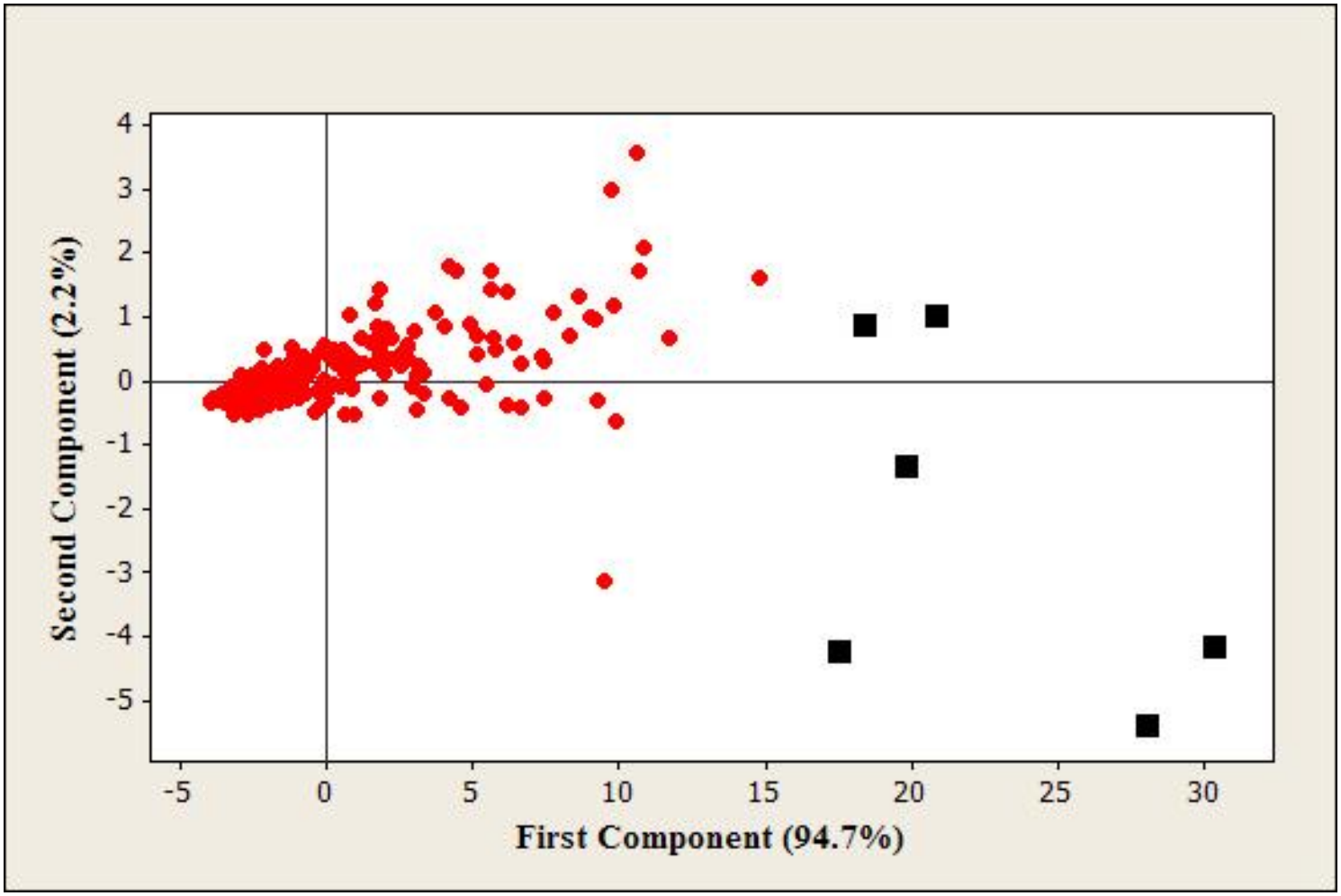


Figure 4.5. Principal component analysis plot for the 283 features. The first principal component accounts for 94.7% of variation and the second component explained only 2.2 % of variation. The squares represent features that account for most of the variability across the genotypes.

Table 4.2. Normalized mean signal intensities of the five features diagnostic for *E. purpurea* and the six features chosen by PCA across the 27 genotypes.

Line	I9 [a]	A8 [a]	O2 [a]	G16 [a]	B17 [a]	J8 [a]	M2 [b]	B15 [b]	N6 [b]	A2 [b]	C2 [b]
pur 669	151.34	169.33	116.40	149.31	140.39	97.20	31.25	7.63	20.79	10.89	22.66
pur "White purpurea"	192.97	187.28	124.37	135.26	135.41	98.00	31.33	8.36	16.79	21.02	28.68
pur 313	198.66	219.55	166.22	136.86	94.23	101.35	36.49	10.42	20.28	31.18	29.76
pur "purpurea"	209.94	243.24	161.43	116.41	126.90	116.54	43.94	8.89	29.43	28.36	31.55
pur 307	84.79	94.92	91.67	100.29	117.59	109.30	55.38	7.79	29.08	30.20	35.16
pur plot 10009	241.51	194.13	176.30	109.53	104.54	110.60	49.18	7.09	30.49	32.98	35.89
pal plot 5	138.55	136.21	93.49	114.15	112.92	96.74	76.33	23.34	48.07	35.04	41.27
ang plot 9	202.66	186.95	140.22	127.23	133.72	104.28	63.95	33.61	36.46	35.16	41.62
px - px 321	166.46	180.79	95.19	119.74	147.38	115.67	60.76	34.81	38.83	36.09	44.64
E. tennesseensis	156.71	165.03	117.53	109.60	100.86	102.26	67.33	18.93	33.41	42.21	45.24
pal 275	199.04	209.54	116.03	136.11	127.21	109.27	65.13	38.72	39.21	36.71	47.05
px - px 301	151.19	160.92	88.69	123.42	143.27	130.93	62.47	23.91	33.49	42.23	47.14
px - neg 264	152.85	168.71	93.56	109.34	132.47	111.90	68.93	36.22	47.89	41.00	47.22
ang - ang 318	183.71	157.32	105.66	139.21	113.35	103.47	56.99	34.14	36.16	38.54	48.05
px - neg 265	166.58	163.31	101.58	119.66	129.81	110.59	61.97	36.75	42.57	41.43	48.63

Line	I9 [a]	A8 [a]	O2 [a]	G16 [a]	B17 [a]	J8 [a]	M2 [b]	B15 [b]	N6 [b]	A2 [b]	C2 [b]
px - neg 263	141.89	186.09	79.50	110.74	140.81	100.72	67.55	41.91	40.46	42.91	49.33
pal 293	201.80	216.30	114.15	138.44	105.96	121.24	65.14	36.29	35.24	35.55	50.63
hyb 294	95.85	107.20	60.41	93.07	100.00	86.66	69.44	28.58	42.40	47.33	52.08
pal 315	139.24	162.91	88.69	96.89	121.92	130.29	82.06	43.42	58.30	50.05	52.43
ang - str 320	142.00	158.58	93.47	94.45	81.61	95.16	74.39	28.10	40.06	50.61	54.30
pal "Hula dancer"	122.28	127.91	92.74	100.78	115.51	115.58	80.14	33.63	44.64	56.49	54.72
ang – ang 285	106.89	131.26	76.73	103.86	116.53	93.70	80.86	36.51	52.12	49.98	55.06
ang 267	118.82	166.44	101.69	93.52	115.87	115.12	78.51	44.08	62.67	53.41	55.80
pal 290	131.24	126.91	74.86	103.34	105.52	114.03	81.16	34.23	47.85	55.33	56.62
ang – ang 272	123.98	140.19	82.71	108.60	108.87	95.16	70.90	37.51	39.88	45.87	57.56
pal 296	138.93	138.88	80.54	94.30	130.63	104.73	92.85	36.17	51.70	54.39	58.71
ang - str 266	137.95	175.01	109.14	109.05	121.91	116.83	93.29	40.91	53.86	55.27	60.13
Mean	**155.48**	**165.74**	**105.30**	**114.56**	**119.45**	**107.68**	**65.47**	**28.59**	**39.71**	**40.75**	**46.37**
Variance	**1439.61**	**1185.39**	**806.00**	**265.17**	**263.58**	**118.84**	**271.36**	**155.08**	**126.97**	**122.97**	**101.32**

[a] Features that were chosen by PCA.

[b] Features that were found to have low signal strength for *E. purpurea*.

Furthermore, the magnitude of the variance for the full set of features was examined across the 27 genotypes. Five species-specific features were identified which were not previously detected by PCA since they had low means across the fingerprints **(Table 4.2)**. These results confirm the results obtained in Chapter 2, where PCA was only able to detect the features with high variance and high mean. These five features presented low signal strength for all *E. purpurea* lines analyzed, thus it differentiated *E. purpurea* from the other fingerprinted species.

Pearson bivariate correlation performed on the 11 features chosen above (6 PCA and five features diagnostic for *E. purpurea*), indicated that there are positive significant correlations among the features **(Appendix 14)**. Feature I9 was found to be highly correlated to features O2 ($r = 0.83$, $P<0.01$) and A8 ($r = 0.84$, $P<0.01$) and feature M2 was found to be highly correlated to features N6 ($r = 0.90$, $P<0.01$), A2 ($r = 0.90$, $P<0.01$) and C2 ($r = 0.92$, $P<0.01$). Therefore, O2, A8, N6, A2 and C2 were eliminated from the set of polymorphic features since I9 and M2 could explain most of the variation found in them. Based on the above analysis, only 6 features were selected since they were the most discriminatory and each had unique variation patterns.

A second hierarchical cluster analysis was performed with these 6 features **(Figure 4.6)**. A comparison between this new dendrogram with the original constructed with the full set of features **(Figure 4.4)** revealed that the *E. purpurea* lines were not clearly differentiated from the other species in the new dendrogram as was found in the original. In addition, the *E. purpurea* line PI631307 (pur307) is not clustering with the

other *E. purpurea* genotypes. Consequently, it may be inferred that these six features are not the most useful features for the fingerprinting of these 27 genotypes.

Finally, based on the fact that the first two dendrograms **(Figure 4.4 and 4.6)** generated with the 27 lines were not able to elucidate the relationship between the species a third dendrogram was generated by merging the data of all the lines belonging to the same species. This average of the normalized mean signal intensity was performed for all the 283 features. This dendrogram **(Figure 4.7)** clearly differentiated *E. purpurea* from the other four species, as was found in the original dendrogram. However, *E. angustifolia* and *E. pallida* clustered together and were closely related to *E. paradoxa* which was not clearly shown in the two previous dendrograms.

4.3.3 Correlation of the molecular profile with metabolic profiling

No significant correlations were found between the signal strength of the six features used to construct the second dendrogram with the relative abundance of 43 lipophilic metabolites in roots (Wu et al., 2009). Consequently, the signal strength of the each of the 283 features was used for the correlations with each of the 43 lipophilic metabolites across the 19 lines shared by the two studies. Positive correlation was found between the signal strength of feature H9, L2 and M8 with the relative content of 2,4 diene alkamides and Chen alkamide. In addition, the signal strength of I18 and F15 had positive significant correlations with the relative content of monoene alkamides. Signal of F15 was also found correlated to ketone 24 **(Table 4.3)**.

The most significant correlations were found between the signal strength of feature H9 and the content of chen alkamide (r =0.92; P<0.01), amide 3 (r =0.87; P<0.01) and amide 7 (r =0.87; P<0.01). The signal strength of H9 and the content of the amides were the highest for PI631307 and PI631313 lines **(Figure 4.8)**, which are the only two *E. purpurea* lines used in both studies. The same results were also found for the positive correlations between the signal strength of H9, L2 and M8 with the content of amide 2, 3, 7 and chen alkamide. Furthermore, the signal strength of I18 and content of the amides 14 and 16 were at the highest for the *E. angustifolia* var. *angustifolia* lines **(Figure 4.9)**. This indicates that the signal strength of features H9, L2 and M8 have a similar pattern of variations as the relative content of the amides 2, 3, 7 and chen amide in the two *E. purpurea* lines analyzed. Similarly, the signal strength of feature I18 has a similar pattern of variation as the content of amide 14 and 16 in the *E. angustifolia* var. *angustifolia* lines analyzed.

4.3.4. The sequence identity of the most interesting features

The amplification products of the six features used to construct the second dendrogram were sequenced along with the amplification products of the five features whose signal strength was found to be positively correlated to the content of lipophilic metabolites. Three features (G16, J8 and M2) had significant alignments with putative retrotransposon sequences while feature I9 had a good match to a retrotransposon called *RIRE1* **(Table 4.4)**. It could be seen that features G16 and J8 significantly matched to the same retrotransposon locus. After performing sequence alignment (blastN) it was found that the two sequences partially overlapped by 93 bp. However, both features were found to have different patterns of variation as it was found after correlation

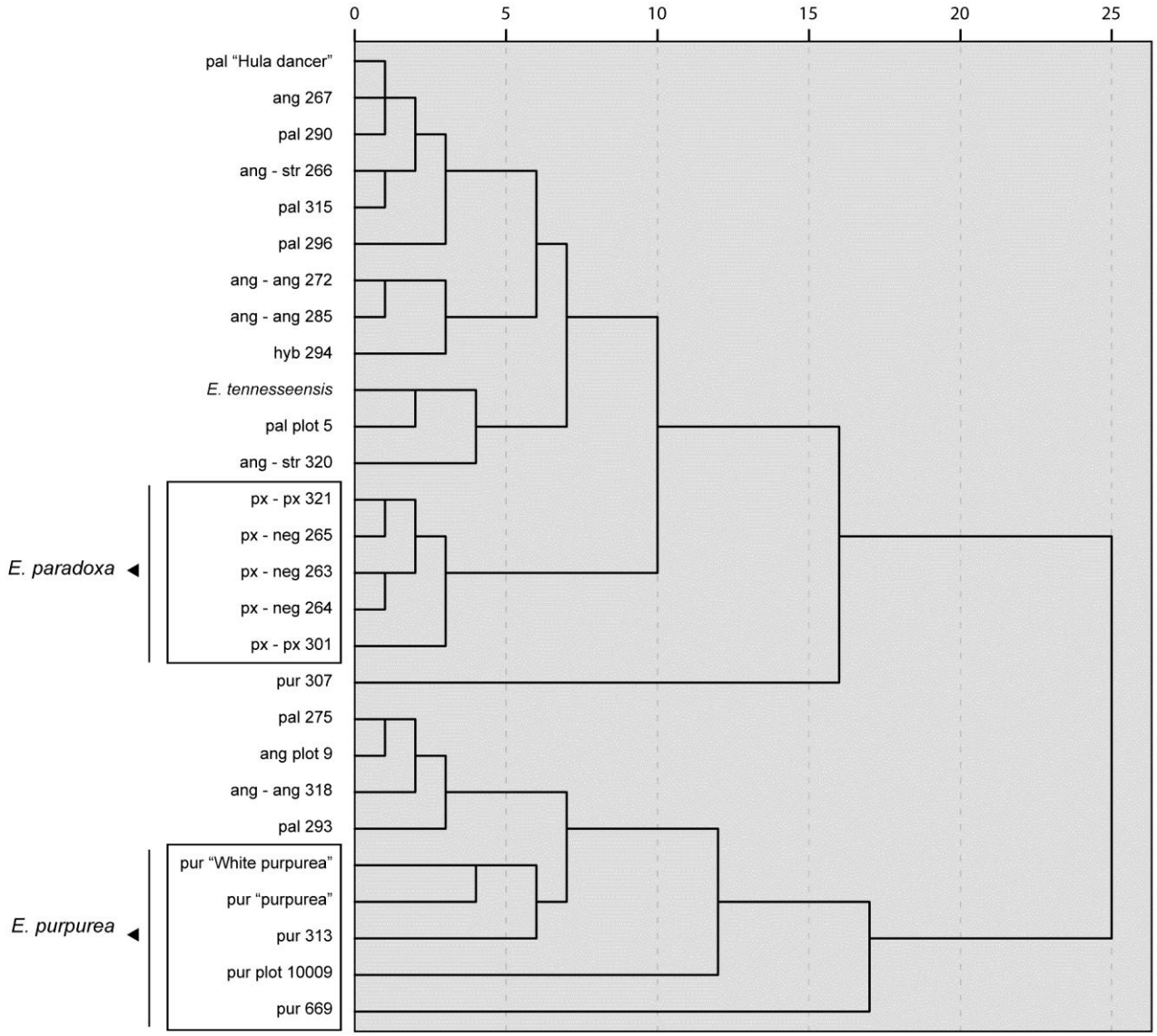


Figure 4.6. Dissimilarity dendrogram (Squared Euclidian distance, between groups linkage) for the SDA hybridization patterns of the 27 genotypes using only the six most discriminatory features. The steps of the dendrogram show the combined clusters and the values of the distance coefficients at each step; the values have been rescaled to numbers between 0 and 25, preserving the ratio of the distances between the steps. The equivalents of the abbreviated names used for each of the lines are shown in Table 4.1.

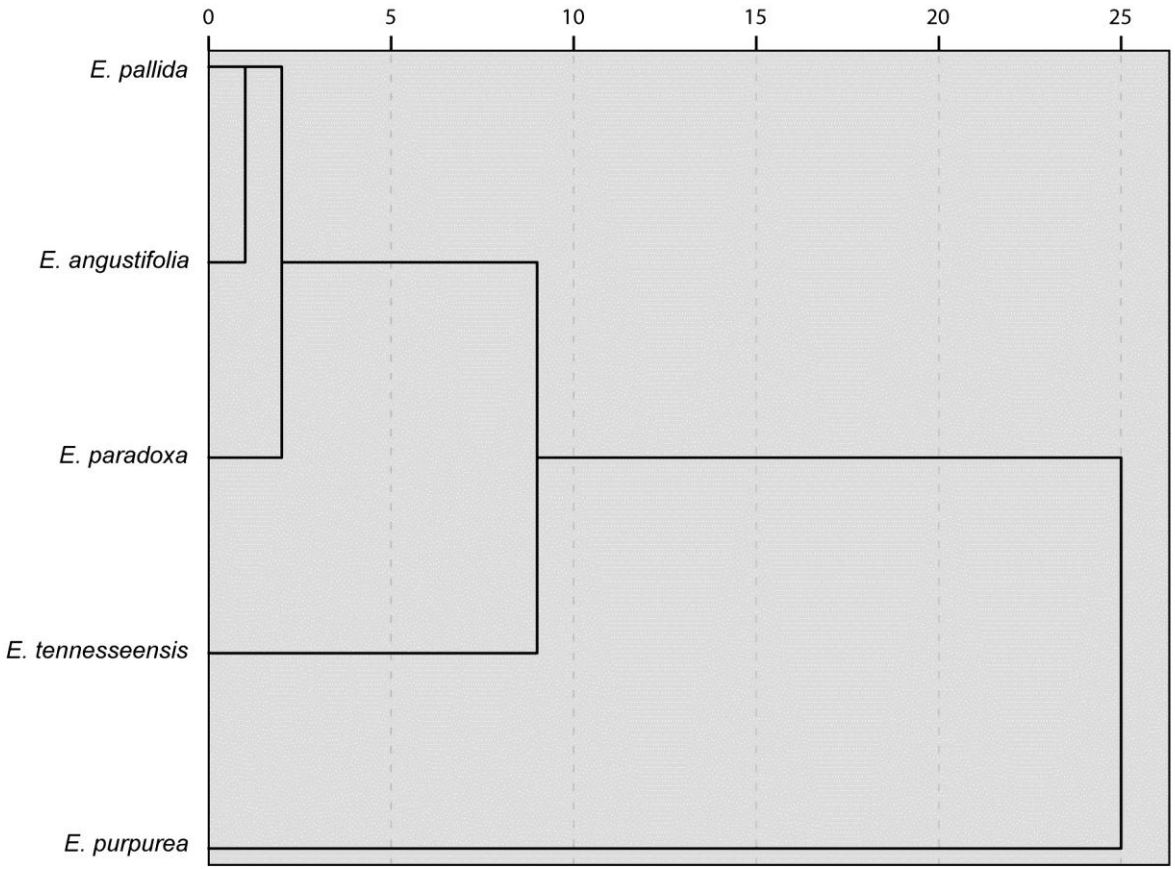


Figure 4.7. Dissimilarity dendrogram (Squared Euclidian distance, between groups linkage) generated by merging the data of all the accessions belonging to the same species. This average of the normalized mean signal intensity was performed for all the 283 features. The steps of the dendrogram show the combined clusters and the values of the distance coefficients at each step; the values have been rescaled to numbers between 0 and 25, preserving the ratio of the distances between the steps.

Table 4.3. Significant correlations among the signal of each of the 283 features and the relative abundance of 43 lipophilic metabolites.

Signal / Compound	H9	L2	M8	I18	F15
Amide 1 [a]		**0.82** **0.00**			
Amide 2 [a]	**0.65** **0.00**	**0.81** **0.00**			
Amide 3 [a]	**0.87** **0.00**		**0.74** **0.00**		
Amide 4 [a]					
Amide 5				**0.54*** **0.02**	
Amide 7 [a]	**0.86** **0.00**	**0.79** **0.00**	**0.75** **0.00**		
Amide 8					
Amide 9				**0.57*** **0.02**	
Amide 10 [a]				**0.59** **0.00**	
Amide 11 [a]				**0.48*** **0.04**	
Amide 12 [b]				**0.49*** **0.03**	
Amide 13 [b]				**0.59** **0.00**	
Amide 14 [b]				**0.71** **0.00**	
Amide 15					**0.79** **0.00**
Amide 16 [b]				**0.75** **0.00**	
Amide 17 [b]				**0.59** **0.00**	
Chen alkamide	**0.92** **0.00**	**0.71** **0.00**	**0.70** **0.00**		
Ketone 22			**-0.56*** **0.02**		
Ketone 24			**-0.61** **0.00**		**0.56*** **0.02**

[a] 2,4-diene alkamides

[b] Monoene alkamides

** Correlation is significant at the 0.01 level

*Correlation is significant at the 0.05 level

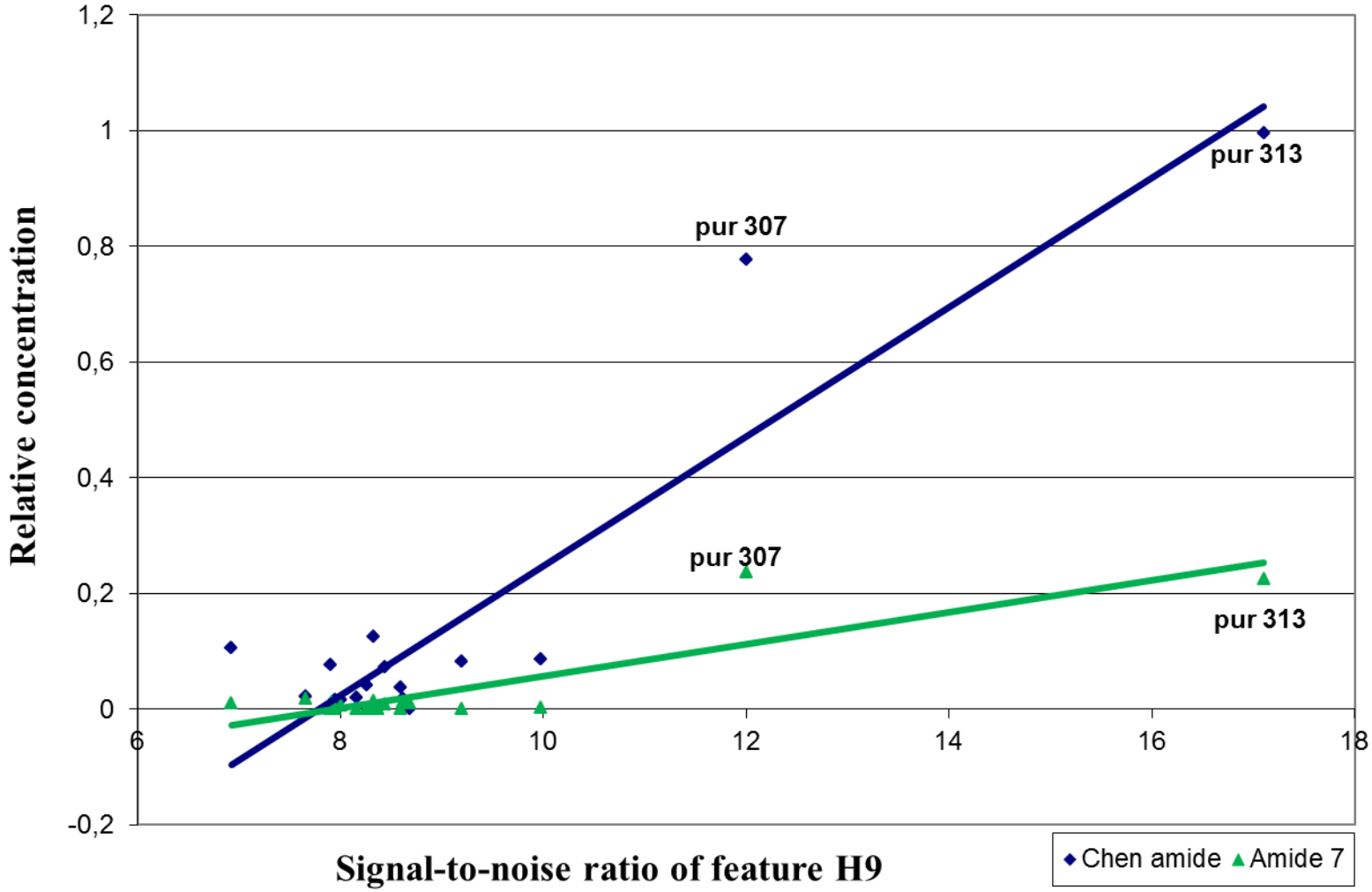


Figure 4.8. Significant correlation among the signal strength of feature H9 and the relative contents of Chen alkamide and amide 7.

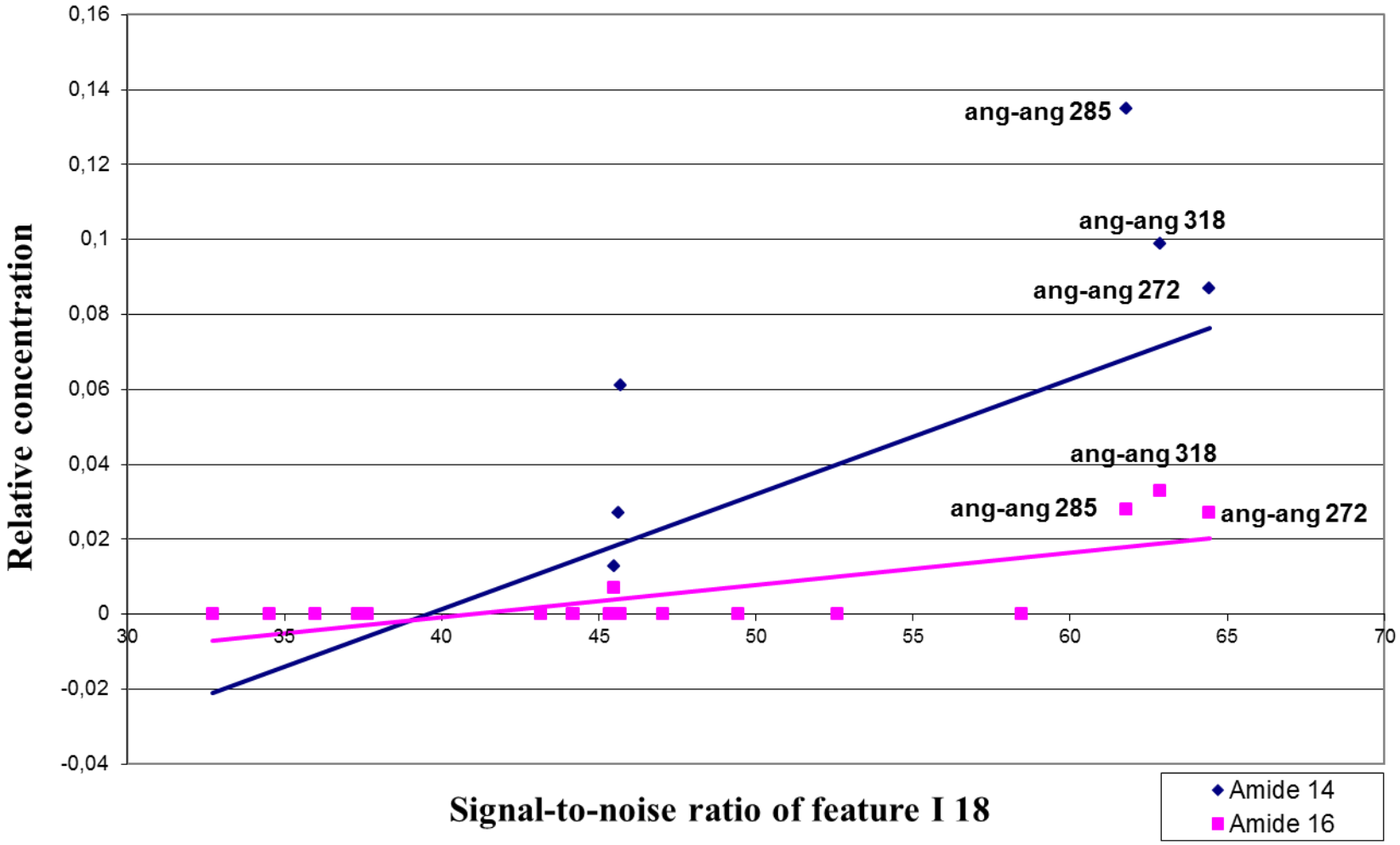


Figure 4.9. Significant correlation among the signal strength of feature I18 and the relative contents of amide 14 and 16.

Table 4.4. Predicted locus/function of the 11 sequenced SDA features using blastN program through National Centre of Biotechnology Information (www.ncbi.nlm.nih.gov). Showing the best match as the putative identity for each sequence. E-value regarded as significant if < 1e-10. NA indicates the absence of significant data.

Feature ID	Length (bp)	Matching database entry	Putative identity	E Value
B15 [b]	252	-	No hits	NA
B17 [a]	344	EL419699.1	*Helianthus ciliaris* uncharacterized cDNA sequence	2e-11
F15 [c]	744	EU362851.1	Ambrosia asymptomatic virus 2 UKM-2007 isolate 05TGP00321.Bad4 ORF1-2 gene, parcial cds	6e-54
G16 [a]	328	FJ791047.1	*Helianthus annuus* retrotransposon HA7, complete sequence	3e-21
H9 [c]	249	-	No hits	NA
I9 [a]	550	D85597.1	*Oryza australiensis* retrotransposon RIRE1 DNA	6e-08
I18 [c]	447	-	No hits	NA
J8 [a]	300	FJ791047.1	*Helianthus annuus* retrotransposon HA7, complete sequence	2e-11
L2 [c]	643	JN021935.1	*Helianthus annuus* cutivar HA383 clone BAC 0516M24, complete sequence	4e-43
M2 [b]	829	GQ367282.1	*Helianthus petiolaris* isolate 94XPET9 retrotransposon Ty3/gypsy-like reverse transcriptase-like gene, partial sequence	1e-95
M8 [c]	454	-	No hits	NA

[a] Features that were chosen by PCA.

[b] Features that were found to have low signal strength for *E. purpurea*.

[c] Features which signal strength correlated significantly with the content of lipophilic metabolites.

bivariate where no significant correlation was found between these two features (r = -0.004, P >0.05). Furthermore, feature L2 matched to the sequence of a bacterial artificial chromosome (BAC) clone. The sequence of this feature in the BAC clone was found between a *copia*- and a *gypsy*-like retrotransposons, however this fragment did not have its own identity. Feature F15 significantly matched to an ORF 1-2 gene, and feature B17 corresponded to an uncharacterized cDNA sequence. The other four remaining features were not recognized as known DNA sequences or proteins **(Table 4.4)**.

The efficiency of the suppression PCR effect was also estimated by observing if each of the sequences possessed different adaptors at their ends. Out of eleven features sequenced, ten features had sequences with both adaptors at their ends **(Appendix 15)**. Only the sequence of feature J8 did not have any of the adaptors, which may have resulted in the poor sequence quality at the beginning and end of the read. None of the features had the same adaptor sequences at both ends, which implies that the suppression PCR may have been efficient at removing these sequences.

4.4 DISCUSSION

4.4.1 Subtraction efficiency

The subtraction technique was able to eliminate about 97% of common DNA sequences between the tester (*Echinacea*) and the driver pool (non-angiosperms, angiosperms but

excluding the Asteraceae). This subtraction efficiency was higher than the one obtained for the *Salvia* study (88%), detailed in **Section 2.4.1** and identical to the one obtained for the prototype SDA for angiosperms (Jayasinghe et al., 2007), where 12 (3%) features were found to be positively hybridized to the driver DNA. Therefore, modifications to the subtraction protocol subsequent to the *Salvia* study increased the efficiency of this process considerably. The factors contributing to this improvement are discussed below.

Salvia subtraction was performed with the PCR-Select™ cDNA Subtraction Kit (Clontech), however for *Echinacea* subtraction the PCR-Select™ Bacterial Genome Subtraction Kit (Clontech) was used. Both kits are suitable for plant DNA subtractions, however thirty-fold excess of driver (1:30 tester:driver ratio) is the recommended ratio for cDNA subtraction compared to sixty-fold excess (1:60 tester:driver ratio) used for the Bacterial Kit. These ratios were optimized by Clontech for cDNA or bacterial DNA subtraction, however they were not optimized for subtraction of plant genomic DNA. The (1:30) ratio was used for the subtraction performed by Jayasinghe et al. (2007) which effectively eliminated the common sequences between tester (angiosperm) and driver (non-angiosperm). However, the *Echinacea* and *Salvia* subtractions were more stringent that the one performed by Jayasinghe et al. (2007) since the subtraction needed to be performed to the genus level. It may be possible that a higher concentration of driver DNA may be needed to remove all homologous sequences from the tester pools. Mathematical models for suppression subtractive hybridization have shown that by increasing the excess ratios the likelihood for enriching for highly specific tester sequences is increased (Gadgil et al., 2002). Therefore, the sixty-fold excess of driver

added during the *Echinacea* subtraction may have removed most of the sequences that were homologous between the *Echinacea* and driver pools. The thirty-fold excess of driver used in the *Salvia* study may have not been sufficient to eliminate such sequences.

The sixty-fold excess used for *Echinacea* subtraction was the only modification made to the established SSH protocol described in **Section 2.2.2.2**. Other modifications could have been performed, for instance the presence of features with the same adaptor at both sites in the *Salvia* array **(Section 2.4.1)** suggested that suppression PCR conditions needed to be optimized to avoid amplification of type b molecules. However, since the tester:driver ratio was increased, a decision was made to leave the PCR parameters constant. Therefore, this increase in the excess ratios may possibly be the contributing factor for the increased efficiency in the *Echinacea* subtraction. Consequently, future subtractions made at the genus level should be performed using a 1:60 tester:driver ratio.

4.4.2 Genetic relationships among twenty-seven *Echinacea* lines

The species within each cluster in each of the hierarchical analyses generated were labeled on the basis of McGregor's classification. This classification was based on morphological traits and chromosome numbers performed on a wide sampling of wild populations, where all *Echinacea* were subsequently assigned to nine species and four varieties (McGregor, 1968). However, the results presented in this study are not in agreement with this classification, for three primary reasons.

Firstly, *E. purpurea* was clearly differentiated from the other four species in the hierarchical cluster analyses performed with full set of features **(Figure 4.4)** and the one were the lines belonging to the same species were merged **(Figure 4.7)**. **Figure 4.7** shows a distance threshold of more than 20 between *E. purpurea* and the cluster that contain the other three species. This result agrees with the conclusions from Binns's et al. (2002a) study where two major divergent taxa within *Echinacea* were found. In this classification *E. purpurea* is the only member in the subgenus *Echinacea* and the subgenus *Pallida* included all other taxa.

Secondly, there is no clear differentiation between the *E. pallida* and *E. angustifolia* lines. As shown in **Figure 4.4**, the seven lines from each of these two species did not cluster as expected according to the species or varieties from which they belonged (sensu McGregor). For instance, the *E. pallida* (PI631293 and PI631275) and *E. angustifolia* lines (PI 631318 and Plot 9) were found in cluster 2 which was clearly differentiated from cluster 3 and 4 that contained the other lines of these two species. This result is more in agreement with Binns's et al. (2002a) classification, where *E. angustifolia* is a variety of *E. pallida* which contains five varieties (*E. pallida* (Nutt.) Nutt. var. *angustifolia* (DC.) Cronq, *E. pallida* (Nutt.) Nutt. var. *pallida*, *E. pallida* (Nutt.) Nutt. var. *sanguinea* (Nutt.) Gandhi & R.D. Thomas, *E. pallida* (Nutt.) Nutt. var. *simulata* (McGregor) Binns, B.R. Baum, & Arnason and *E. pallida* var. *tennesseensis* (Beadle) Binns B.R.Baum, & Arnason).

Thirdly, the current results could not support the classification into varieties by either McGregor (1968) or Binns et al. (2002a). McGregor recognized four varieties (*E.*

angustifolia DC. var. *angustifolia*, *E. angustifolia* DC. var. *strigosa* McGregor, *E. paradoxa* (Norton) Britton var. *neglecta* McGregor and *E. paradoxa* (Norton) Britton var. *paradoxa*) of which only the lines belonging to *E. paradoxa* var. *neglecta* clustered together **(Figure 4.4)**. Binns recognized eight varieties (five of *E. pallida*, *E. atrorubens* (Nutt.) Nutt. var. *atrorubens*, *E. atrorubens* (Nutt.) Nutt. var. *paradoxa* (J.B. Norton) Cronq. and *E. atrorubens* (Nutt.) Nutt. var. *neglecta* (McGregor) Binns, B.R. Baum, & Arnason), however it was not possible to support entirely this classification since not all the taxa could be included in this study (*E. atrorubens, E. laevigata*, *E. sanguinea* and E. *simulata*) due to quarantine restrictions.

Although some of the results explained above support the classification by Binns et al. (2002a), the SDA profile could not unequivocally support the division by either four species or eight varieties due to the smaller number of species used. Therefore, further studies including all species are needed in order to elucidate the genetic relationships for all *Echinacea* species. To date, it has not been possible to reconstruct the genetic and evolutionary relationships of this genus due to limitations in the population sampling and in the use of techniques such as AFLP and RAPD which have to make the assumption that co-migrating fragments are homologous, limiting its applications for phylogenetic analyses (Kapteyn et al., 2002; Kim et al., 2004; Mechanda et al., 2004a). In addition, the use of chloroplast and nuclear loci, which are commonly used for phylogenetic studies, were unable to resolve the species level relationships due to the low levels of molecular divergence found in these loci (Flagel et al., 2008). SDA offers a good alternative for nuclear DNA-based phylogenetic analysis; however the results from this study provide an incomplete assessment of the phylogenetic relationship of the

genus since not all the species were analyzed. Future studies could concentrate on the missing species using the SDA developed with the aim to provide a more comprehensive phylogenetic and evolutionary analysis of the genus.

4.4.3 The SDA for authentication purposes

The SDA was capable of fingerprinting genotypes that were not used in its construction **(Figure 4.4)**. For instance it was possible to fingerprint *E. angustifolia* (ang plot 9), *E. pallida* (pal plot 5), *E. purpurea* (pur plot 10009) and a hybrid of *E. paradoxa* var. *paradoxa* with *E. pallida* (hyb 294). The *E. purpurea* plot 10009 grouped with the other lines from the same specie in the original dendrogram **(Figure 4.4)**, which indicates that the array may possibly be used to identify unknown samples of *E. purpurea*. The lines of *E. angustifolia*, *E. pallida* and the hybrid grouped with lines from the same or closely related species, however if an unknown sample is fingerprinted the only two species that could be unambiguously identified will be *E. paradoxa* and *E. purpurea*, since the SDA did not obtain a clear distinction between *E. angustifolia* and *E. pallida*. Therefore, the use of this SDA for authentication purposes may be limited. In addition, the use of a hybridization technique for authentication purposes would be more time consuming and expensive than established PCR based techniques. Furthermore, a PCR based technique may be not be developed as easily as in *Salvia* since the six highly polymorphic features were not the main features that established the hierarchical subdivision and species-specific features were not identified for each of the species. Previous studies have already identified specific RAPD markers (Nieri et al., 2003; Wolf et al., 1999) and specific AFLP markers (Russi et al., 2009) for the three most commonly used medicinal species (*E. pallida*, *E. angustifolia* and *E. purpurea*). Moreover, Mechanda et al.

(2004a) found that an individual plant could be identified to species with 10 AFLP bands from a single primer set. Therefore, AFLP and RAPD markers may be more efficient for the genotyping and identification of *Echinacea* species relative to the SDA.

4.4.4 Correlations between the genetic and chemical profiles

The significant positive correlations found for the hybridization profile of H9, L2 and M8 with the contents of 2,4 diene alkamides in the lines or accessions analyzed could be attributed to the fact that 2,4 dienoic acid unit is present in *E. purpurea* in higher amounts (Binns et al., 2002b) and signal strength of these features is relatively higher for *E. purpurea* lines. Therefore, these three features could be good markers for *E. purpurea*, however they could not be consider as potential markers for 2,4 diene alkamides since the signal strength and the relative abundance of the amides do not share a similar pattern of variation for all other species. The same problem was found for feature I18, where the signal strength of I18 have a similar pattern of variation as the relative content of amides 14 and 15 only for *E. angustifolia* var. *angustifolia* and not with the other species. Consequently, the significant correlations found could indicate that these loci may potentially be species-specific markers rather than markers linked to genes responsible for the production of these bioactive compounds.

It is important to note that even though this study used the same accessions as Wu et al. (2009), and sourced these accessions from the same germplasm collection, different plants were used for each study. Previous studies have found that populations and cultivars of *Echinacea* are genetically heterogeneous (Chuang et al., 2010b; Kapteyn et al., 2002). Therefore, the fact that these two studies were performed on different plants

may be a possible reason for the different patterns of variation among signal strength of the features and the relative content of the lipophilic metabolites. Furthermore, when the dendrogram constructed with the SDA data **(Figure 4.4)** and the one generated from 43 lipophilic metabolites (Wu et al., 2009) were compared, the only similarity found is that the lines of *E. pallida* did not cluster together in both genetic and chemical profiles. Future studies should employ the same plants to perform chemical and molecular analysis in order to identify molecular markers associated not only with the production of lipophilic alkamides but also to caffeic acid derivates and ketoalken/ynes which are important bioactive compounds found in *Echinacea* (Binns et al., 2002b).

Previous studies have found DNA molecular markers useful for predicting the phytochemical concentration of *Echinacea* plants. AFLP DNA fingerprints were found to be statistically significant as predictors of cichoric acid and dodeca-2E, 4E, 8Z, 10E/Z-tetraenoic acid isobutyl amide (amide 8 and 9) in cultivated *E. purpurea* and some related wild species (Baum et al., 2001), also RAPD markers were able to predict polyphenol content in aerial parts of *E. purpurea* (Chen et al., 2009a). However, to date no study has performed a correlation analysis that includes all species of *Echinacea*. Binns et al. (2002b) determined by a phytochemical variation study that the best taxonomical markers for species delimitation in *Echinacea* root materials were the amides 1, 2, 3 and 7 together with cichoric acid and ketoalkene 24 (pentadeca-8Z,13Z-dien-11-yn-2-one). Future studies could perform chemical and molecular profiles with all the species in order to find if species-specific markers could also be associated to the production of bioactive compounds, since the abundance of the compounds varies greatly depending on the species (Wu et al., 2009).

4.4.5 Identity of the most interesting features

Out of eleven features sequenced four corresponded to known retrotransposon loci. Retrotransposons are mobile genetic elements which can be classified in two clearly separate groups, the long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons (Schulman et al., 2004). Features G16 and J8 matched to the same database entry, *Helianthus annuus* retrotransposon HA7 which is a putative LTRs (Vukich et al., 2009). Feature I9 and M2 also matched to LTR retrotransposons. Feature I9 had a good match to a retrotransposon named *RIRE1* (for Rice Retroelement) (Noma et al., 1997), while M2 significantly matched to a *Ty3/gypsy*-like LTR retrotransposon (Ungerer et al., 2009). LTR- retrotransposons have been found to be more prevalent in plant genomes (can comprise about 50% of the nuclear DNA) and have been found to play a major role in the expansion of the genome size (Bennetzen et al., 2005). For instance, it has been found that *RIRE1* caused an increase in size of about 11 Mb in *Oryza australiensis* (Noma et al., 1997). The high abundance of LTR retrotransposons in the genome, their ubiquitous nature and their activity in creating genomic diversity by stably integrating large DNA segments into dispersed chromosomal loci, make this group of retrotransposons ideal for development as molecular markers (Schulman et al., 2004). Previous studies have found that it is possible to use retrotransposons for fingerprinting cultivated rice species (Kang and Kang, 2007), to obtain genomic diversity patterns of *Pisum* (Jing et al., 2007) and to elucidate the evolutionary events of three *Helianthus* hybrid species independently derived from two parental species (Ungerer et al., 2009). The results obtained in the present study suggest that LTR retrotransposons are highly polymorphic in *Echinacea*; therefore the four loci that matched to known retrotransposons have the potential to

become retrotransposon-based molecular markers useful for fingerprinting and studying diversity patterns in *Echinacea*.

It is important to note that among the eleven features sequenced there was no match to any chloroplast loci as was the case in the *Salvia* study. In contrast, the majority of matches were to retrotransposons. There could be several reasons that may explain the differences found between the sequences identified on both arrays. For instance, the features sequenced were the ones found to be the more polymorphic within the genus; therefore these loci could have been present in the array however they were not identified as the highly polymorphic ones. Several studies have used polymorphic regions of the chloroplast genome for fingerprinting *Salvia* (Takano and Okada, 2010; Walker and Sytsma, 2007; Walker et al., 2004), however in *Echinacea* these commonly used chloroplast loci do not display much divergence (Flagel et al., 2008). In addition, it may be possible that these sequences may have been present in the enriched genus-specific sequences obtained from the subtraction process prior to cloning, however since only about 300 positive clones were picked these sequences may have not been cloned, thus they may have not been present in the array. Another possible reason is that these sequences were removed during the subtraction process since they were homologous with the driver. For instance, if chloroplast sequences in *Echinacea* are highly homologous with the one from the driver, then these sequences should have been removed during subtraction. Although it is not possible to determine that the above reasons were in fact the cause of the different sequences identified on both arrays, the sequence of all features in both arrays by future studies could clarify if the differences

are due to the process of array construction or to the differences between the two genera analyzed.

4.5 CONCLUSIONS

The efficient enrichment of specific sequences during subtraction (97%) made it possible to obtain a set of unique sequences for *Echinacea*. The *Echinacea*-SDA constructed successfully fingerprinted the twenty-seven *Echinacea* lines and differentiated *E. purpurea, E. paradoxa* from the other species. However no clear differentiation was observed between the *E. pallida* and *E. angustifolia* lines. These results provided support for the classification sensu Binns et al. (2002a), however due to the smaller number of species used in this study it was not possible to unequivocally support the division of *Echinacea* into four species and eight varieties as proposed by this morphometric classification.

Moreover, significant correlations were found among the genetic and chemical profiles; however the different patterns of variation among signal strength of the features and the relative content of the lipophilic metabolites could probably imply more a correlation with specific species than a correlation with the content of the biocompounds. Therefore, future association studies should use the same individual plants to perform chemical and molecular analysis in order to confirm positive correlations.

Although the SDA was enriched with *Echinacea* specific sequences, the only species-specific features identified were for *E. purpurea* and *E. angustifolia* var. *angustifolia*. However, four retrotransposon sequences were identified to be polymorphic among the 27 genotypes, which imply they could be potential genotyping markers useful for fingerprinting and studying diversity patterns in *Echinacea*.

CHAPTER 5

Conclusions

5.1 INTRODUCTION

Although *Salvia* and *Echinacea* have important medicinal and commercial value, correct species identification is challenging. In both of these genera, misidentification among closely related species is common due to their morphological similarity. Considering that SDA does not require previous DNA sequence information and has been shown in previous studies to be capable of genotyping medicinal plants to the clade and family level, this technique was employed to develop two specific SDAs able to fingerprint *Salvia* and *Echinacea*.

Although these two studies were performed using the same technique of subtraction, array construction, hybridization, scanning and data analysis, the main findings were very different for each genus. This chapter provides conclusions and compares the main findings of each study. **Figures 5.1 and 5.2** summarize the main findings of the *Salvia* and *Echinacea* studies respectively.

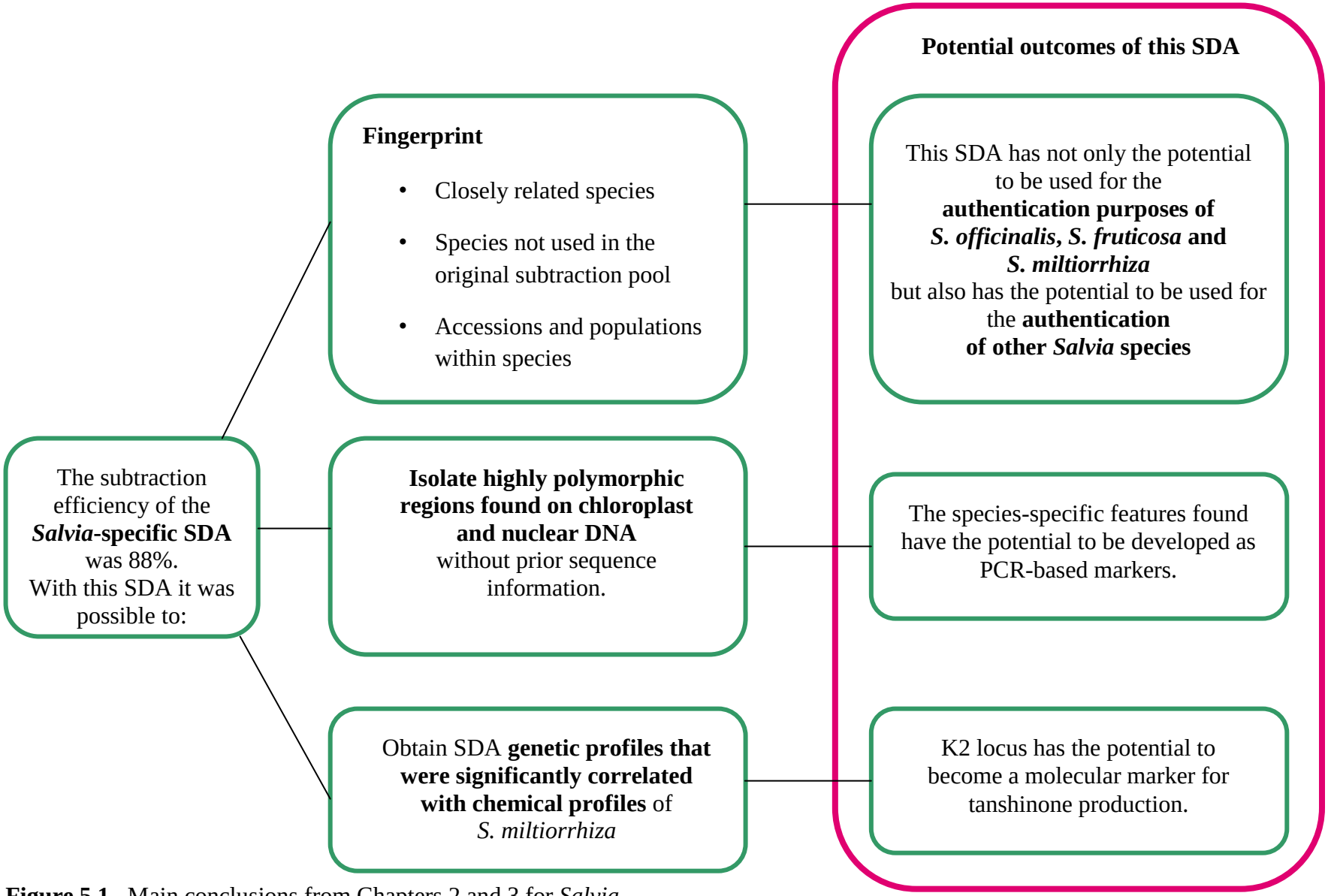


Figure 5.1. Main conclusions from Chapters 2 and 3 for *Salvia*.

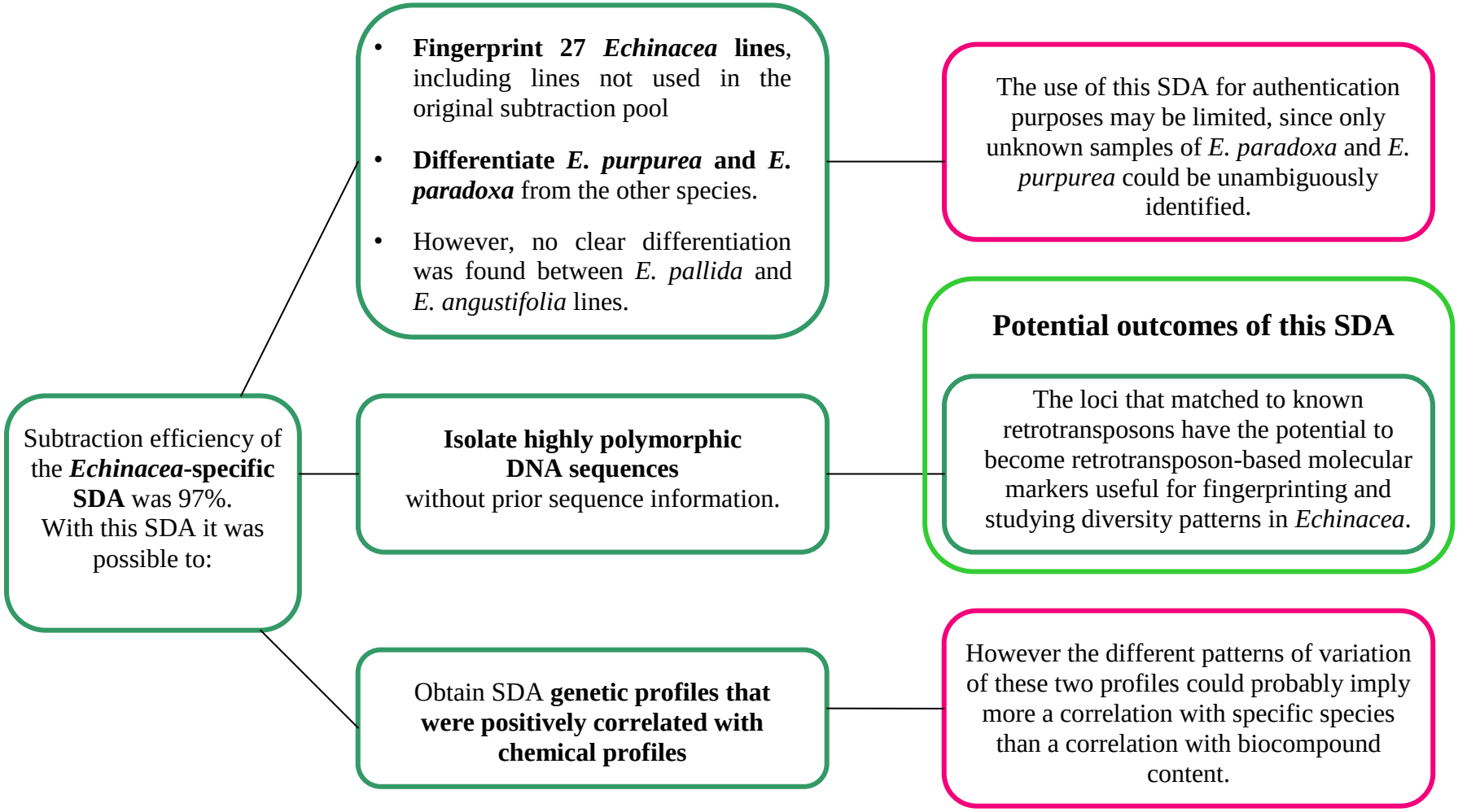


Figure 5.2. Main conclusions from Chapter 4 for *Echinacea*.

5.2 SUBTRACTION EFFICIENCY AND LEVEL OF POLYMORPHISM

The suppression subtractive hybridization technique (SSH) was used to enrich selectively each SDA with highly specific DNA sequences of the genera *Salvia* and *Echinacea*. The subtraction efficiency obtained for the *Echinacea*-SDA was higher than for the *Salvia*-SDA, possibly due to an increase in the tester:driver ratios from 1:30 to 1:60. However, the fact that the subtraction efficiency was higher for the *Echinacea*-SDA does not imply that the level of polymorphism in this array was higher. The subtraction and further cloning isolated sequences that were specific for the taxa under study but many of them could be monomorphic across the species. For instance, in order to differentiate among lines and species of *Echinacea*, more stringent hybridization conditions were needed (increase in hybridization temperature and more stringent washes), when compared to the conditions used for *Salvia* **(Section 4.2.3)**. This higher stringency used for *Echinacea* may indicate that the level of polymorphism in the *Echinacea*-SDA was lower than for the *Salvia*-SDA. However, the level of polymorphism in the arrays could not be measured since the data obtained in this study was scored based on the signal intensity (signal–to-noise ratio) instead of using presence/absence (dominant scoring), as explained in **Section 2.4.2**. It is known from previous array studies that the level of polymorphism will depend directly on the level of genetic diversity available within the taxa used to develop the array (Gupta et al., 2008). The level of genetic diversity among *Echinacea* species has been found to be relatively low in some nuclear and chloroplast loci (Flagel et al., 2008; Urbatsch et al., 2000). This lower level of genetic diversity expected in *Echinacea* compared to *Salvia* may possibly explain why the discrimination between species was more challenging for the *Echinacea* genus than for the *Salvia* genus and may possibly indicate a lower level

of polymorphism in the *Echinacea*-SDA even though the subtraction efficiency was higher for this array.

One analysis that was not included in this study was an analisis to determine the molecular basis of SDA polymorphisms. To date, no previous SDA study has determined which kind of polymorphism may be detected by this technique. Similar techniques such as DArT generally detect single base pair changes (SNPs) and insertion /deletion/rearrangements polymorphisms in restriction enzyme sites **(Section 1.4.4.1)**. Future studies could, for instance, develop specific primers for the most polymorphic features found in this study in order to amplify these loci in species, accessions and populations of these two genera. Then the amplification products could be isolated, sequenced and aligned in order to reveal which microstructural changes are responsible for the segregation of these SDA features, elucidating in this way the basis of the SDA polymorphism.

5.3 THE SDA FOR AUTHENTICATION PURPOSES

The SDAs were able to genotype species, populations and accessions of the taxa under study. The *Salvia*-SDA was able to accurately genotype up to the species and population level, and the *Echinacea*-SDA was able to genotype twenty-seven lines. In addition, both arrays were able to fingerprint species not used in the construction of the original subtraction pool which indicates that the SDA can be constructed using a representative pool of individuals instead of all members of the group under study. This is a practical advantage for authentication of a genus like *Salvia*, which is a broad genus

(approximately 1000 species) that exhibits a wide range of morphological and ecological variation. The *Salvia*-SDA was constructed with only ten species and it has the potential to be used not only for authentication purposes of *S. officinalis, S. fruticosa* and *S. miltiorrhiza* but also has the potential to be used for the authentication of other *Salvia* species **(Section 2.4.3)**. However, the use of the *Echinacea*-SDA for authentication purposes may be limited, since this array only differentiated *E. purpurea* and *E. paradoxa* from the other species and no clear differentiation was observed between the *E. pallida* and *E. angustifolia* lines **(Section 4.3.2)**. This result is in agreement with Binns's et al. (2002a) classification, where *E. angustifolia* is a variety of *E. pallida*, however the results of the present study could not completely support this morphometric classification due to the smaller number of species used in this study. Future studies could clarify if this close relationship found between the *E. pallida* and *E. angustifolia* lines is due to high genetic similarities between them or to a low discriminatory power of the *Echinacea*-SDA. For instance, specific primers could be developed for some of the most interesting features in order to amplify and compare the sequences from these two species and find if there are polymorphisms that could not be detected by the SDA and that are able to differentiate these two species. Future studies are then needed in order to validate the use of the *Echinacea*-SDA for authentication purposes.

It is important to note that if these two arrays want to be employed for authentication purposes they would need to be able to fingerprint possible species out the genera that could be used as adulterants or substitutes. However, species out of these two genera were not fingerprinted and used as out-groups in the present study. There are two main

reasons that explain the absence of out-groups in the study. Firstly these were genera specific arrays, which imply that genomic sequences from species outside of the genera would show low or no hybridization signal. For instance, *Tanacetum parthenium* and *Eupatorium perfoliatum* were hybridized to the *Echinacea*-SDA, however in order to obtain a minimal signal 3 times more biotin-labeled sample was needed. Similarly, S*cutellaria lateriflora* was hybridized to the *Salvia*-SDA, however lower signals were obtained after increasing the biotin-labeled sample 3 times. Secondly, the previous angiosperm-SDA (Jayasinghe et al., 2007) could complement the authentication of *Salvia* and *Echinacea* by fingerprinting possible contaminants or adulterants that do not belong to these two genera. Previous studies have already shown that this array is capable of fingerprinting to the species level with minor exceptions (Jayasinghe et al., 2009). Furthermore, the discriminatory power of this angiosperm-SDA could be significantly increased by the development of two clade-specific SDAs, one for the Asterids and the other for the Rosids. Our RMIT research group recently constructed the Asterid and the Rosids clade-specific SDA using the suppression subtractive hybridization employed in this study and then validated these two arrays. The Asterid-SDA has already been used to fingerprint 25 Asterid species representing 20 families and 9 orders within this clade (Mantri et al., 2011). Therefore, the combination of the angiosperm, clade and genera specific arrays into one, may allow a fingerprinting of medicinal herbs from clade level down to species and even accession or variety level, allowing the identification of the correct species or variety and possible adulterants.

5.4 THE SDA AS A DISCOVERY TOOL FOR POLYMORPHIC SEQUENCES

The broad subtraction approach conducted to produce the SDA was able to isolate highly polymorphic regions found in nuclear DNA. For instance, several unknown and uncharacterized polymorphic sequences were identified in both arrays which are almost certainly part of nuclear DNA, since if they were part of chloroplast or mitochondrial DNA they would have matched to an accession in GenBank **(Section 2.4.5)**. Furthermore, polymorphic retrotransposon sequences were identified in the *Echinacea*-SDA **(Section 4.4.5)**. These nuclear sequences identified have several advantages, for example, they are biparentally inherited which makes them useful for identification of hybrids. Additionally, they could be associated or linked to a gene responsible for an important agronomical trait, since these genes are usually found in the nuclear genome. Furthermore, the nuclear sequences identified in this study were not only polymorphic across the species analyzed but also they were highly specific for each of the genera studied. Therefore, these sequences have the potential to become PCR-based markers and be employed for species authentication, marker assisted selection (such as K2 locus) and phylogenetic analyses where additional genomic regions to chloroplast and ITS loci may provide improved phylogenetic resolution. Previous studies have also performed selection of nuclear loci using sequence databases as a framework; however this approach is dependent on the phylogenetic proximity of the taxa under study to the species available in sequence databases (Alvarez et al., 2008). Therefore, the broad subtraction conducted to produce the SDA is a good alternative to identify new polymorphic loci for authentication of medicinal herbs that do not have closely related species with available genomic libraries.

5.5 THE SDA AS A POTENTIAL TOOL FOR DETECTING MARKERS ASSOCIATED WITH IMPORTANT AGRONOMICAL TRAITS

The SDA also could be used to discover potential markers associated with important agronomical traits. The hybridization patterns obtained with the SDA for *S. miltiorrhiza* and *Echinacea* significantly correlated with chemical profiles obtained in previous studies. However after comparing the patterns of variation among signal strength of the features and the relative content of the metabolites a good correlation was only found between the K2 locus and the content of tanshinones in *S. miltiorrhiza* **(Section 3.4.4)**, while for *Echinacea* the different patterns of variation found could imply more of a correlation with specific species **(Section 4.4.4)**. The different patterns of variation in *Echinacea* may be attributed in part to the fact that molecular and chemical studies were performed on different plants. Previous studies have found that populations and cultivars of *Echinacea* are genetically heterogeneous which may imply that using different plants from the same population for chemical or molecular fingerprinting could produce misleading interpretations since each plant could have a different profile. However, a good correlation was obtained in the *Salvia* study even though different plants were also used. Another possible reason is that while for *Salvia* the correlations were performed only for *S. miltiorrhiza* and *S. sinica* which are species that produced tanshinones with similar concentrations, for *Echinacea* the correlations were performed for four species where the abundance of the compounds varies greatly depending on the species. For instance, most of the previous studies that have found DNA molecular markers useful for predicting the phytochemical concentration in *Echinacea* have only been performed in *E. purpurea*. To date no study has performed a correlation analysis that includes all species of *Echinacea*. Therefore, in order to confirm the positive

correlations found in this study future association studies should employ the same individual plants to perform both the chemical and molecular analyses. In addition, future studies should perform these correlations on individual species as well as with all the species used in the study.

5.6 THE SDA FOR PHYLOGENETIC ANALYSES

The primary purpose of the genus-specific SDAs constructed in this study was authentication; however, the results also highlight the possibility of using SDA for phylogenetic analysis. For instance, the hierarchical cluster dendrogram constructed on the *Salvia* species revealed genetic relationships consistent with geographical origins **(Section 2.4.4)**. However, there were some clear differences between previous phylogenetic studies performed in this genus and the one obtained with the SDA as explained in **Section 2.4.4**. The previous angiosperm-SDA developed by Jayasinghe et al. (2007) was capable of differentiating accurately six angiosperm clades, which implied the SDA technique, could be used to establish phylogenetic relationships. The main difference in the development of the angiosperm and *Salvia* SDA was in the representation of the substraction pool (tester pool). For the construction of the angiosperm-SDA the tester pool was equally represented with all the angiosperm clades, however during the construction of the SDA, the subtraction pool was enriched with *S. miltiorrhiza*, *S. sinica* and *S. officinalis*. Therefore, the SDA could be over-represented with sequences from these three species, and as a result the phylogenetic analyses obtained from this array could be biased in terms of the distances given across the species and major clusters. Based on this result, for the construction of *Echinacea*-SDA

the tester pool ideally should have had all species equally represented; however due to quarantine restrictions only five species out of nine (sensu McGregor) were used. Although not all species were used in the construction of the *Echinacea*-SDA it may be possible that this SDA as well as the retrotransposon sequences identified could be used for phylogenetic analyses since the species are so closely related. In conclusion, the enrichment of the SDA for species-specific sequences may be detrimental for phylogenetic purposes since it could introduce a bias in terms of the distances given in the dendrogram; however it could be an advantage for authentication purposes of those specific species as it was shown in the *Salvia* study.

5.7 CONCLUDING REMARKS

This study constructed two genera-specific SDAs that have the potential to authenticate species, populations and accessions with the advantage that it may also be possible to authenticate species not included in the construction of the array. The ability to fingerprint species not used in the development of the SDA is mainly due to the presence of polymorphic sequences specific for the genus studied which were isolated by the broad subtraction approach conducted to produce these two arrays. Additionally, some of these highly polymorphic sequences isolated have also the potential to become molecular markers that may be used to predict the content of bioactive compounds. For instance, K2 locus, which hybridization patterns significantly correlated with the chemical profiles for tanshinones in *S. miltiorrhiza*, has the potential to become a molecular marker for tanshinone production that may be used in the standardization of these active components in *Salvia* and in the selection of optimum genotypes. Future

studies should however validate the use of K2 locus as a marker. Similarly it will be important to validate the use of these two SDAs for identification of unknown and adulterated samples in order to validate the use of these SDAs for authentication purposes. In conclusion, the results showed that the SDA technique may have more than one potential use, for instance it may be used for authentication purposes (Niu et al., 2011b), or also it may be employed for phylogenetic analyses (Jayasinghe et al., 2009) as shown in the present and previous studies. This study particularly highlighted another use of this technique, which is the use of the SDA as a discovery tool for potential polymorphic markers that could be employed in species identification and marker assisted selection.

BIBLIOGRAPHY

Adinolfi, B., Chicca, A., Martinotti, E., Breschi, M.C., Nieri, P., 2007. Sequence characterized amplified region (SCAR) analysis on DNA from the three medicinal *Echinacea* species. Fitoterapia 78, 43-45.

Agarwal, M., Shrivastava, N., Padh, H., 2008. Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep. 27, 617-631.

Akbari, M., Wenzl, P., Caig, V., Carling, J., Xia, L., Yang, S., Uszynski, G., Mohler, V., Lehmensiek, A., Kuchel, H., Hayden, M.J., Howes, N., Sharp, P., Vaughan, P., Rathmell, B., Huttner, E., Kilian, A., 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet 113, 1409-1420.

Alarcon-Aguilar, F.J., Roman-Ramos, R., Flores-Saenz, J.L., Aguirre-Garcia, F., 2002. Investigation on the hypoglycaemic effects of extracts of four Mexican medicinal plants in normal and alloxan-diabetic mice. Phytother Res 16, 383-386.

Altamirano-Dimas, M., Hudson, J.B., Cochrane, D., Nelson, C., Arnason, J.T., 2007. Modulation of immune response gene expression by *Echinacea* extracts: results of a gene array analysis. Can J Physiol Pharmacol. 85, 1091-1098.

Alvarez, I., Costa, A., Feliner, G.N., 2008. Selecting single-copy nuclear genes for plant phylogenetics: a preliminary analysis for the Senecioneae (Asteraceae). J Mol Evol. 66, 276-291.

Alvarez, I., Wendel, J.F., 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29, 417-434.

Arnason, J.T., Binns, S.E., Baum, B.R., 2002. Phytochemical diversity and biological activity in *Echinacea* phytomedicines: Challenges to quality control and germplasm improvement. In: Meskin, M.S. (Ed.), Phytochemicals in Nutrition and Health. CRC Press.

Barrett, B., 2003. Medicinal properties of *Echinacea*: a critical review. Phytomedicine 10, 66-86.

Barthelson, R.A., Sundareshan, P., Galbraith, D.W., Woosley, R.L., 2006. Development of a comprehensive detection method for medicinal and toxic plant species. Am J Bot 93, 566-574.

Baum, B.R., Binns, S.E., Arnason, J.T., 2004. Taxonomic History and Revision of the Genus *Echinacea*. In: Miller, S.C. (Ed.), *Echinacea*. CRC Press.

Baum, B.R., Binns, S.E., Arnason, J.T., 2006. Integrating recent knowledge about the genus *Echinacea*: morphology, molecular systematics and phytochemistry. Herbal Gram 72, 32-46.

Baum, B.R., Mechanda, S., Livesey, J.F., Binns, S.E., Arnason, J.T., 2001. Predicting quantitative phytochemical markers in single *Echinacea* plants or clones from their DNA fingerprints. Phytochemistry 56, 543-549.

Bennetzen, J.L., Ma, J., Devos, K.M., 2005. Mechanisms of recent genome size variation in flowering plants. Ann Bot 95, 127-132.

Bergeron, C., Livesey, J.F., Awang, D.V.C., Arnason, J.T., Rana, J., Baum, B.R., Letchamo, W., 2000. A quantitative HPLC method for the quality assurance of *Echinacea* Products on the North American market. Phytochem Anal 11, 207-215.

Binns, S.E., Baum, B.R., Arnason, J.T., 2002a. A Taxonomic Revision of *Echinacea* (Asteraceae: Heliantheae). Syst. Bot. 27, 610-632.

Binns, S.E., Livesey, J.F., Arnason, J.T., Baum, B.R., 2002b. Phytochemical variation in *Echinacea* from roots and flowerheads of wild and cultivated populations. J Agric Food Chem 50, 3673-3687.

Blumenthal, M., Urbatsch, L.E., 2006. *Echinacea* taxonomy - is the re-classification of the genus warranted? Herbal Gram 72, 30-31,80.

Böszörményi, A., Héthelyi, E., Farkas, A., Horváth, G., Papp, N., Lemberkovics, E., Szoke, E., 2009. Chemical and genetic relationships among sage (*Salvia officinalis* L.) cultivars and Judean sage (*Salvia judaica* Boiss.). J. Agric. Food Chem . 57, 4663-4667.

Bremer, B., Bremer, K.r., Chase, M., Fay, M., Reveal, J., Soltis, D., Soltis, P., Stevens, P., 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Linn. Soc. 161, 105-121.

Bruna, S., Giovannini, A., Benedetti, L.D., Principato, M.C., Ruffoni, B., 2006. Molecular Analysis of *Salvia* Spp . through RAPD markers. ISHS Acta Hortic. 723, 157-160.

Bruni, I., De Mattia, F., Galimberti, A., Galasso, G., Banfi, E., Casiraghi, M., Labra, M., 2010. Identification of poisonous plants by DNA barcoding approach. Int. J. Legal. Med. 124, 595-603.

Cahill, J.P., 2004. Genetic diversity among varieties of Chia (*Salvia hispanica* L.). Genet. Resour. Crop Evol. 51, 773-781.

Cai, Z., Lee, F.S.C., Wang, X.R., Yu2, W.J., 2002. A capsule review of recent studies on the application of mass spectrometry in the analysis of Chinese medicinal herbs. J Mass Spectrom 37, 1013 -1024.

Canter, P.H., Thomas, H., Ernst, E., 2005. Bringing medicinal plants into cultivation: opportunities and challenges for biotechnology. Trends Biotechnol 23, 180-185.

Cao, J., Wei, Y.J., Qi, L.W., Li, P., Qian, Z.M., Luo, H.W., Chen, J., Zhao, J., 2008. Determination of fifteen bioactive components in Radix et Rhizoma Salviae Miltiorrhizae by high-performance liquid chromatography with ultraviolet and mass spectrometric detection. Biomed Chromatogr 22, 164-172.

Carles, M., Cheung, M.K., Moganti, S., Dong, T.T., Tsim, K.W., Ip, N.Y., Sucher, N.J., 2005. A DNA microarray for the authentication of toxic traditional Chinese medicinal plants. Planta Med 71, 580-584.

Chan, K., 2003. Some aspects of toxic contaminants in herbal medicines. Chemosphere 52, 1361-1371.

Chase, M.W., Salamin, N., Wilkinson, M., Dunwell, J.M., Kesanakurthi, R.P., Haidar, N., Savolainen, V., 2005. Land plants and DNA barcodes: short-term and long-term goals. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 360, 1889-1895.

Chavan, P., Joshi, K., Patwardhan, B., 2006. DNA microarrays in herbal drug research. Evid Based Complement Alternat Med 3, 447-457.

Chen, C.L., Chuang, S.J., Chen, J.J., Sung, J.M., 2009a. Using RAPD markers to predict polyphenol content in aerial parts of *Echinacea purpurea* plants. J Sci Food Agri 89, 2137-2143.

Chen, D., Zhao, S.S., Leung, K.S., 2009b. Improved quality assessment of proprietary Chinese medicines based on multi-chemical class fingerprinting. J Sep Sci 32, 2892-2902.

Chen, S.L., 2010. The strategies on medicinal plant genome sequencing project. The 9th Meeting of the Consortium for Globalization of Chinese Medicine, Hong Kong on August 23-25, 2010.

Chiou, S.J., Yen, J.H., Fang, C.L., Chen, H.L., Lin, T.Y., 2007. Authentication of medicinal herbs using PCR-amplified ITS2 with specific primers. Planta Med 73, 1421-1426.

Chuang, S., Chen, C., Chen, J., Sung, J., 2010a. Using morphological diagnosis and molecular markers to assess the clonal fidelity of micropropagated *Echinacea purpurea* regenerants. Biologia Plantarum 54, 539-542.

Chuang, S.J., Chen, C.L., Chen, J.J., Sung, J.M., 2010b. Using bulked AFLP analysis to assess genetic diversity in *Echinacea* species. Sci. Hortic. 124, 400-404.

Claßen-Bockhoff, R., 2007. Floral Construction and Pollination Biology in the Lamiaceae. Ann. Bot. 100, 359-360.

Coram, T.E., Pang, E.C.K., 2005. Isolation and analysis of candidate ascochyta blight defence genes in chickpea. Part I. Generation and analysis of an expressed sequence tag (EST) library. Physiol. Mol. Plant. Pathol. 66, 192-200.

Deng, K.J., Zhang, Y., Xiong, B.Q., Peng, J.H., Zhang, T., Zhao, X.N., Ren, Z.L., 2009. Identification, characterization and utilization of simple sequence repeat markers derived from *Salvia miltiorrhiza* expressed sequence tags. Yao xue xue bao = Acta pharmaceutica Sinica 44, 1165-1172.

Diatchenko, L., Lau, Y.F., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E.D., Siebert, P.D., 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci U S A 93, 6025-6030.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 1-15.

Drasar, P., Moravcova, J., 2004. Recent advances in analysis of Chinese medical plants and traditional medicines. J Chromatogr B. 812, 3-21.

Dudai, N., Lewinsohn, E., Larkov, O., Katzir, I., Ravid, U., Chaimovitsh, D., Sa'adi, D., Putievsky, E., 1999. Dynamics of yield components and essential oil production in a commercial hybrid sage (*Salvia officinalis* x *Salvia fruticosa* cv. Newe Ya'ar no. 4). J Agric Food Chem 47, 4341-4345.

Duncan, B.D., Isaac, G., 1994. Ferns and allied plants of Victoria, Tasmania and South Australia. Melbourne University Press in association with Monash University.

Echeverrigaray, S., Agostini, G., 2006. Genetic relationships between commercial cultivars and Brazilian accessions of *Salvia officinalis* L. based on RAPD markers. Rev. Bras. Pl. Med. 8, 13-17.

Fan, X.-H., Chenga, Y.-Y., Ye, Z.-L., Lin, R.-C., Qian, Z.-Z., 2006. Multiple chromatographic fingerprinting and its application to the quality control of herbal medicines. Anal. Chim. Acta 555 217–224.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M., Barrett, S.C.H., 2008. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS One 3, e2802.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C.H., Newmaster, S.G., Hajibabaei, M., Husband, B.C., 2009. Are plant species inherently harder to discriminate than animal species using DNA barcoding markers?. Mol. Ecol. Resour. 9, 130-139.

Flagel, L.E., Rapp, R.A., Grover, C.E., Widrlechner, M.P., Hawkins, J., Grafenberg, J.L., Alvarez, I., Chung, G.Y., Wendel, J.F., 2008. Phylogenetic, morphological, and chemotaxonomic incongruence in the North American endemic genus *Echinacea*. Am J Bot 95, 756-765.

Frederich, M., Jansen, C., de Tullio, P., Tits, M., Demoulin, V., Angenot, L., 2009. Metabolomic analysis of Echinacea spp. by 1H nuclear magnetic resonance spectrometry and multivariate data analysis technique. Phytochem Anal 21, 61-65.

Gadgil, C., Rink, A., Beattie, C., Hu, W.S., 2002. A mathematical model for suppression subtractive hybridization. Comp Funct Genom 3, 405-422. .

Gao, H., Wang, Z., Li, Y., Qian, Z., 2011. Overview of the quality standard research of traditional Chinese medicine. Frontiers of medicine 5, 195-202.

Gnavi, G., Bertea, C.M., Maffei, M.E., 2009. PCR, sequencing and PCR–RFLP of the 5S-rRNA-NTS region as a tool for the DNA fingerprinting of medicinal and aromatic plants. Flavour Fragr J 25, 132-137.

Guo, B.L., Feng, Y.X., Zhao, Y.J., 2002. Review of germplasm resources studies on *Salvia miltiorrhiza*. China journal of Chinese materia medica 27, 492-495.

Guo, L., Fuscoe, J.C., Fu, P., Mei, N., 2009. Application of DNA Microarray in Studies of Herbal Dietary Supplements. Handbook of Systems Toxicology. John Wiley & Sons, Ltd.

Gupta, P.K., Rustgi, S., Mir, R.R., 2008. Array-based high-throughput DNA markers for crop improvement. Heredity 101, 5-18.

Han, J.P., Liu, C., Li, M.H., Shi, L.C., Song, J.Y., Yao, H., Pang, X.H., Chen, S.L., 2010. Relationship between DNA barcoding and chemical classification of *Salvia* medicinal herbs. Chinese Herbal Medicines 2, 16-29.

He, C.-e., Wei, J., Jin, Y., Chen, S., 2010. Bioactive components of the roots of *Salvia miltiorrhizae*: Changes related to harvest time and germplasm line. Ind Crops Prod 32, 313-317.

Hebert, P.D., Gregory, T.R., 2005. The promise of DNA barcoding for taxonomy. Syst. Biol. 54, 852-859.

Herrera-Ruiz, M., Garcia-Beltran, Y., Mora, S., Diaz-Veliz, G., Viana, G.S., Tortoriello, J., Ramirez, G., 2006. Antidepressant and anxiolytic effects of hydroalcoholic extract from *Salvia elegans*. J Ethnopharmacol 107, 53-58.

Heubl, G., 2010. New aspects of DNA-based authentication of Chinese medicinal plants by molecular biological techniques. Planta Med 76, 1963-1974.

Hou, C.C., Chen, C.H., Yang, N.S., Chen, Y.P., Lo, C.P., Wang, S.Y., Tien, Y.J., Tsai, P.W., Shyur, L.F., 2010. Comparative metabolomics approach coupled with cell- and gene-based assays for species classification and anti-inflammatory bioactivity validation of *Echinacea* plants. J. Nutr. Biochem. 21, 1045-1059.

Hu, P., Liang, Q.L., Luo, G.A., Zhao, Z.Z., Jiang, Z.H., 2005. Multi-component HPLC fingerprinting of *Radix Salviae Miltiorrhizae* and its LC-MS-MS identification. Chem Pharm Bull (Tokyo) 53, 677-683.

Hudson, J., Altamirano, M., 2006. The application of DNA micro-arrays (gene arrays) to the study of herbal medicines. J Ethnopharmacol 108, 2-15.

Jaccoud, D., Peng, K., Feinstein, D., Kilian, A., 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29, E25.

James, K.E., Schneider, H., Ansell, S.W., Evers, M., Robba, L., Uszynski, G., Pedersen, N., Newton, A.E., Russell, S.J., Vogel, J.C., Kilian, A., 2008. Diversity arrays technology (DArT) for pan-genomic evolutionary studies of non-model organisms. PLoS One 3, e1682.

Jayasinghe, R., Hai Niu, L., Coram, T.E., Kong, S., Kaganovitch, J., Xue, C.C.L., Li, C.G., Pang, E.C.K., 2009. Effectiveness of an Innovative Prototype Subtracted Diversity Array (SDA) for Fingerprinting Plant Species of Medicinal Importance. Planta Med. 75, 1180-1185.

Jayasinghe, R., Kong, S., Coram, T.E., Kaganovitch, J., Xue, C.C., Li, C.G., Pang, E.C., 2007. Construction and validation of a prototype microarray for efficient and high-throughput genotyping of angiosperms. Plant Biotechnol. J. 5, 282-289.

Jin, L., 2001. Pollination factors in *Salvia miltiorrhiza* Bge. . Journal of Capital Normal University 02.

Jing, R., Johnson, R., Seres, A., Kiss, G., Ambrose, M.J., Knox, M.R., Ellis, T.H., Flavell, A.J., 2007. Gene-based sequence diversity analysis of field pea (Pisum). Genetics 177, 2263-2275.

Joshi, K., Chavan, P., Warude, D., Patwardhan, B., 2004. Molecular markers in herbal drug technology. Curr Sci India 87.

Kamatou, G.P., Makunga, N.P., Ramogola, W.P., Viljoen, A.M., 2008. South African Salvia species: a review of biological activities and phytochemistry. J Ethnopharmacol 119, 664-672.

Kang, H.W., Kang, K.K., 2007. Genomic characterization of *Oryza* species-specific CACTA-like transposon element and its application for genomic fingerprinting of rice varieties. Mol Breed 21, 283-292.

Kapteyn, J., Goldsbrough, B., Simon, E., 2002. Genetic relationships and diversity of commercially relevant *Echinacea* species. Theor Appl Genet 105, 369-376.

Karaca, M., Ince, A.G., Ay, S.T., Turgut, K., Onus, A.N., 2008. PCR-RFLP and DAMD-PCR genotyping for *Salvia* species. J. Sci. Food Agric. 88, 2508-2516.

Kim, D.H., Heber, D., Still, D.W., 2004. Genetic diversity of *Echinacea* species based upon amplified fragment length polymorphism markers. Genome 47, 102-111.

Kindscher, K., 1989. Ethnobotany of purple coneflower (*Echinacea angustifolia*, Asteraceae) and other *Echinacea* species. Econ. Bot. 43, 498-507.

Kingsley, M.T., Straub, T.M., Call, D.R., Daly, D.S., Wunschel, S.C., Chandler, D.P., 2002. Fingerprinting closely related *Xanthomonas pathovars* with random nonamer oligonucleotide microarrays. Appl. Environ. Microbiol. 68, 6361-6370.

Kiran, U., Khan, S., Mirza, K.J., Ram, M., Abdin, M.Z., 2010. SCAR markers: a potential tool for authentication of herbal drugs. Fitoterapia 81, 969-976.

Kuzma, L., Rozalski, M., Walencka, E., Rozalska, B., Wysokinska, H., 2007. Antimicrobial activity of diterpenoids from hairy roots of *Salvia sclarea* L.: salvipisone as a potential anti-biofilm agent active against antibiotic resistant *Staphylococci*. Phytomedicine 14, 31-35.

Laasonen, M., Wennberg, T., Harmia-Pulkkinen, T., Vuorela, H., 2002. Simultaneous analysis of alkamides and caffeic acid derivatives for the identification of *Echinacea purpurea*, *Echinacea angustifolia*, *Echinacea pallida* and *Parthenium integrifolium* roots. Planta Med 68, 572-574.

Lau, E., Olarte, A., Mantri, N., Li, C.G., Xue, C., Pang, E.C.K., 2011. Fingerprinting provincial varieties of the chinese red sage (*Salvia miltiorrhiza*) and the development of DNA-based markers for bioactive compounds. Plant & Animal Genome XIX Conference, Town & Country Convention Center.San Diego, CA.

Lee, C.Y., Sher, H.F., Chen, H.W., Liu, C.C., Chen, C.H., Lin, C.S., Yang, P.C., Tsay, H.S., Chen, J.J., 2008. Anticancer effects of tanshinone I in human non-small cell lung cancer. Mol. Cancer Ther. 7, 3527-3538.

Lee, H.L., Jansen, R.K., Chumley, T.W., Kim, K.J., 2007. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. Mol. Biol. Evol. 24, 1161-1180.

Leung, P.C., Cheng, K.F., 2008. Good Agricultural Practice (GAP)- Does it ensure a perfect supply of medicinal herbs for research and drug development? International Journal of Applied Research in Natural Products 1, 1-8.

Lezar, S., Myburg, A.A., Berger, D.K., Wingfield, M.J., Wingfield, B.D., 2004. Development and assessment of microarray-based DNA fingerprinting in *Eucalyptus grandis*. Theor Appl Genet 109, 1329-1336.

Li, C.G., Sheng, S.J., Pang, E.C., Marriot, P., May, B., Zhou, S.F., Story, D., Xue, C.C.L., 2009a. Cultivar variations of Australian-grown Danshen (*Salvia miltiorrhiza*): bioactive markers and root yields. Chem. Biodivers. 6, 170-181.

Li, C.G., Sheng, S.J., Pang, E.C., May, B., Xue, C.C., 2009b. HPLC profiles and biomarker contents of Australian-grown *Salvia miltiorrhiza* f. alba roots. Chem Biodivers 6, 1077-1086.

Li, M., Zhang, K.Y., But, P.P., Shaw, P.C., 2011. Forensically informative nucleotide sequencing (FINS) for the authentication of Chinese medicinal materials. Chin Med 6, 42.

Li, M.H., 2008. Investigation of Danshen and related medicinal plants in China. J Ethnopharmacol 120, 419-426.

Li, T.-X., Wang, J.-K., Bai, Y.-F., Lu, Z.-H., 2006. Diversity Suppression-Subtractive Hybridization Array for Profiling Genomic DNA Polymorphisms. J. Integr. Plant Biol. 48, 460-467.

Li, T., Wang, J., Bai, Y., Sun, X., Lu, Z., 2004. A novel method for screening species-specific gDNA probes for species identification. Nucleic Acids Res. 32, pg. e45.

Liu, A.H., Lin, Y.H., Yang, M., Guo, H., Guan, S.H., Sun, J.H., Guo, D.A., 2007. Development of the fingerprints for the quality of the roots of *Salvia miltiorrhiza* and its related preparations by HPLC-DAD and LC-MS(n). J Chromatogr B Analyt Technol Biomed Life Sci 846, 32-41.

Longaray-Delamare, A.P., Moschen-Pistorello, I.T., Artico, L., Atti-Serafini, L., Echeverrigaray, S., 2007. Antibacterial activity of the essential olis of *Salvia officinalis* L. and *Salvia triloba* L, cultivated in south Brazil. Food Chem 100, 603-608.

Lu, Y., Foo, L.Y., 2002. Polyphenolics of *Salvia*--a review. Phytochemistry 59, 117-140.

Ma, H.L., Qin, M.J., Qi, L.W., Wu, G., Shu, P., 2007. Improved quality evaluation of *Radix Salvia miltiorrhiza* through simultaneous quantification of seven major active components by high-performance liquid chromatography and principal component analysis. Biomed Chromatogr.

Mace, E.S., Xia, L., Jordan, D.R., Halloran, K., Parh, D.K., Huttner, E., Wenzl, P., Kilian, A., 2008. DArT markers: diversity analyses and mapping in Sorghum bicolor. BMC Genomics 9, 26.

Mader, E., Lohwasser, U., Börner, A., Novak, J., 2010. Population structures of genebank accessions of *Salvia officinalis* L. (Lamiaceae) revealed by high resolution melting analysis. Biochem. Syst. Ecol. 38, 178-186.

Mantri, N., Olarte, A., Li, C.G., Xue, C., Pang, E.C.K., 2011. Fingerprinting the Asterid species using Subtracted Diversity Array reveals novel species-specific sequences. Plos One. (In Press).

Mazza, G., Cottrell, T., 1999. Volatile components of roots, stems, leaves, and flowers of *Echinacea* species. J Agric Food Chem 47, 3081-3085.

McGregor, R.L., 1968. The taxonomy of the genus *Echinacea* (Compositae). The University of Kansas Science Bulletin 68, 113-142.

Mechanda, S.M., Baum, B.R., Johnson, D.A., Arnason, J.T., 2004a. Analysis of diversity of natural populations and commercial lines of *Echinacea* using AFLP. Can. J. Bot. 82, 461-484.

Mechanda, S.M., Baum, B.R., Johnson, D.A., Arnason, J.T., 2004b. Sequence assessment of comigrating AFLP bands in *Echinacea* - implications for comparative biological studies. Genome 47, 15-25.

Mora, S., Millan, R., Lungenstrass, H., Diaz-Veliz, G., Moran, J.A., Herrera-Ruiz, M., Tortoriello, J., 2006. The hydroalcoholic extract of *Salvia elegans* induces anxiolytic- and antidepressant-like effects in rats. J Ethnopharmacol 106, 76-81.

Nieri, P., Adinolfi, B., Morelli, I., Breschi, M.C., Simoni, G., Martinotti, E., 2003. Genetic characterization of the three medicinal *Echinacea* species using RAPD analysis. Planta Med 69, 685-686.

Niu, L., Mantri, N., Li, C.G., Xue, C., Pang, E., 2011a. Array-based techniques for fingerprinting medicinal herbs. Chin Med 6, 18.

Niu, L., Mantri, N., Li, C.G., Xue, C., Wohlmuth, H., Pang, E.C.K., 2011b. Detection of *Panax quinquefolius* in *Panax ginseng* using 'Subtracted Diversity Array'. J. Sci. Food Agric. 91, 1310-1315.

Noma, K., Nakajima, R., Ohtsubo, H., Ohtsubo, E., 1997. RIRE1, a retrotransposon from wild rice *Oryza australiensis*. Genes Genet. Syst. 72, 131-140.

Obradovic, M., Krajsek, S.S., Dermastia, M., Kreft, S., 2007. A new method for the authentication of plant samples by analyzing fingerprint chromatograms. Phytochem Anal 18, 123-132.

Osowski, S., Rostock, M., Bartsch, H.H., Massing, U., 2000. [Pharmaceutical comparability of different therapeutic *Echinacea* preperations]. Forschende Komplementarmedizin und klassische Naturheilkunde = Research in complementary and natural classical medicine 7, 294-300.

Percival, S.S., 2000. Use of *Echinacea* in medicine. Biochem. Pharmacol. 60, 155-158.

Pound, L.M., Wallwork, M.A.B., Potts, B.M., Sedgley, M., 2002. Self-incompatibility in *Eucalyptus globulus* ssp. *globulus* (Myrtaceae). Aust. J. Bot. 50, 365-372.

Putievsky, E., Ravid, U., Diwan-Rinzler, N., Zohary, D., 1990. Genetic affinities and essential oil composition of *Salvia officinalis* L., *S. fruticosa* Mill., *S. tomentosa* Mill. and their hybrids. Flavour Fragr J 5, 121-123.

Rates, S.M.K., 2001. Plants as source of drugs. Toxicon 39, 603-613.

Reales, A., Rivera, D., Palazon, J.A., Obon, C., 2004. Numerical taxonomy study of *Salvia* sect. *Salvia* (Labiatae). Bot. J. Linn. Soc. 145, 353-371.

Rivera, D., Obon, C., Cano, F., 1994. The botany, history and traditional uses of three-lobes sage (*Salvia fruticosa* Miller) (Labiatae). Econ. Bot. 48, 190-195.

Rogers, S.O., Bendich, A.J., 1987. Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. Plant Mol Biol 9, 509-520.

Rotblatt, M.D., 1999. Herbal medicine: a practical guide to safety and quality assurance. West. J. Med. 171, 172-175.

Russi, L., Moretti, C., Raggi, L., Albertini, E., Falistocco, E., 2009. Identifying commercially relevant *Echinacea* species by AFLP molecular markers. Genome 52, 912-918.

Sahoo, N., Manchikanti, P., Dey, S., 2010. Herbal drugs: standards and regulation. Fitoterapia 81, 462-471.

Savelev, S.U., Okello, E.J., Perry, E.K., 2004. Butyryl- and acetyl-cholinesterase inhibitory activities in essential oils of *Salvia* species and their constituents. Phytother Res 18, 315-324.

Schulman, A.H., Flavell, A.J., Ellis, T.H., 2004. The application of LTR retrotransposons as molecular markers in plants. Methods Mol Biol 260, 145-173.

Sertel, S., Eichhorn, T., Plinkert, P.K., Efferth, T., 2011. [Anticancer activity of *Salvia officinalis* essential oil against HNSCC cell line (UMSCC1).]. Hno.

Shaw, P.C., Wong, K.L., Chan, A.W., Wong, W.C., But, P.P., 2009. Patent applications for using DNA technologies to authenticate medicinal herbal material. Chin Med 4, 21.

Sheng, S., 2007. Cultivation and Quality studies of Danshen (*Salvia miltiorrhiza*) in Australia. School of health Science. RMIT University, Melbourne.

Sheng, S.J., Pang, E.C., Xue, C.C., Li, C.G., 2009. Seasonal variations in bioactive marker contents in Australian-grown *Salvia miltiorrhiza* roots. Chem. Biodivers. 6, 551-560.

Skoula, M., El Hilali, I., Makris, A.M., 1999. Evaluation of the genetic diversity of *Salvia fruticosa* Mill. clones using RAPD markers and comparison with the essential oil profiles. Biochem. Syst. Ecol. 27, 559-568.

Song, J., Yao, H., Li, Y., Li, X., Lin, Y., Liu, C., Han, J., Xie, C., Chen, S., 2009a. Authentication of the family Polygonaceae in Chinese pharmacopoeia by DNA barcoding technique. J Ethnopharmacol 124, 434-439.

Song, Z., Li, X., Wang, H., Wang, J., 2010. Genetic diversity and population structure of *Salvia miltiorrhiza* Bge in China revealed by ISSR and SRAP. Genetica 138, 241-249.

Song, Z.Q., Wang, J.H., Wang, H.G., Zhao, F.J., Hao, L.W., 2009b. Studies of the floral biology, breeding characters of *Salvia miltiorrhiza*. Acta Horticulturae Sinica 06.

Sucher, N.J., Carles, M.C., 2008. Genome-Based Approaches to the Authentication of Medicinal Plants. Planta Med 74, 603,623.

Takano, A., Okada, H., 2010. Phylogenetic relationships among subgenera, species, and varieties of Japanese *Salvia* L. (Lamiaceae) J. Plant Res. 124, 245-252.

Techen, N., Crockett, S.L., Khan, I.A., Scheffler, B.E., 2004. Authentication of medicinal plants using molecular biology techniques to compliment conventional methods. Curr Med Chem 11, 1391-1401.

Topcu, G., 2006. Bioactive triterpenoids from *Salvia* species. J Nat Prod 69, 482-487.

Tsoi, P.Y., Woo, H.S., Wong, M.S., Chen, S.L., Fong, W.F., Xiao, P.G., Yang, M.S., 2003. Genotyping and species identification of *Fritillaria* by DNA chips. Yao Xue Xue Bao 38, 185-190.

Ungerer, M.C., Strakosh, S.C., Stimpson, K.M., 2009. Proliferation of Ty3/*gypsy*-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. BMC biology 7, 40.

Urbatsch, L.E., Baldwin, B.G., Donoghue, M.J., 2000. Phylogeny of the coneflowers and relatives (Heliantheae: Asteraceae) based on nuclear rDNA Internal Transcribed Spacer (ITS) sequences and chlorplast DNA restriction site data. Syst. Bot. 25, 539-565.

Vukich, M., Schulman, A.H., Giordani, T., Natali, L., Kalendar, R., Cavallini, A., 2009. Genetic variability in sunflower (*Helianthus annuus* L.) and in the *Helianthus* genus as assessed by retrotransposon-based molecular markers. Theor Appl Genet 119, 1027-1038.

Walker, J.B., Sytsma, K.J., 2007. Staminal evolution in the genus *Salvia* (Lamiaceae): molecular phylogenetic evidence for multiple origins of the staminal lever. Ann. Bot. 100, 375-391.

Walker, J.B., Sytsma, K.J., Treutlein, J., Wink, M., 2004. *Salvia* (Lamiaceae) is not monophyletic: implications for the systematics, radiation, and ecological specializations of *Salvia* and tribe Mentheae. Am. J. Bot. 91, 1115-1125.

Wang, B., Zhang, Y., Chen, C.B., Li, X.L., Chen, R.Y., Chen, L., 2007. Analysis on genetic diversity of different *Salvia miltiorrhiza* geographical populations in China. China journal of Chinese materia medica 32, 1988-1991.

Wang, C.Y., Chiao, M.T., Yen, P.J., Huang, W.C., Hou, C.C., Chien, S.C., Yeh, K.C., Yang, W.C., Shyur, L.F., Yang, N.S., 2006. Modulatory effects of *Echinacea purpurea* extracts on human dendritic cells: a cell- and gene-based study. Genomics 88, 801-808.

Wang, J.W., Wu, J.Y., 2010. Tanshinone biosynthesis in *Salvia miltiorrhiza* and production in plant tissue cultures. Appl. Microbiol. Biotechnol. 88, 437-449.

Wang, Q., Zhang, B., Lu, Q., 2009. Connserved region amplification polymorphism (CoRAP) a novel marker technique for plant genotyping in *Salvia miltiorrhiza*. Plant Mol Biol Rep 27, 139-143.

Wenzl, P., Carling, J., Kudrna, D., Jaccoud, D., Huttner, E., Kleinhofs, A., Kilian, A., 2004. Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci U S A 101, 9915-9920.

Wenzl, P., Li, H., Carling, J., Zhou, M., Raman, H., Paul, E., Hearnden, P., Maier, C., Xia, L., Caig, V., Ovesna, J., Cakir, M., Poulsen, D., Wang, J., Raman, R., Smith, K.P., Muehlbauer, G.J., Chalmers, K.J., Kleinhofs, A., Huttner, E., Kilian, A., 2006. A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC Genomics 7, 206.

Wills, R.B.H., Perry, N.B., Stuart, D.L., 2004. Factors affecting *Echinacea* quality: Agronomy and processing. In: Miller, S.C. (Ed.), *Echinacea*. CRC Press.

Wittenberg, A.H.J., van der Lee, T., Cayla, C., Kilian, A., Visser, R.G.F., Schouten, H.J., 2005. Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. Mol Genet Genomics 274, 30-39.

Wolf, H.T., Zundorf, I., Winckler, T., Bauer, R., Dingermann, T., 1999. Characterization of *Echinacea* species and detection of possible adulterations by RAPD analysis. Planta Med 65, 773-774.

WorldHealthOrganization, 2008. Traditional medicine Fact sheet N°134 http://www.who.int/mediacentre/factsheets/fs134/en/.

Wu, L., Dixon, P.M., Nikolau, B.J., Kraus, G.A., Widrlechner, M.P., Wurtele, E.S., 2009. Metabolic profiling of *Echinacea* genotypes and a test of alternative taxonomic treatments. Planta Med 75, 178-183.

Xia, L., Peng, K., Yang, S., Wenzl, P., de Vicente, M.C., Fregene, M., Kilian, A., 2005. DArT for high-throughput genotyping of Cassava (*Manihot esculenta*) and its wild relatives. Theor Appl Genet 110, 1092-1098.

Xie, Y., 2006. A high-throughput genomic tool: Diversity array technology complementary for rice genotyping. J Integr Plant Biol 48, 1069-1076.

Xu, H., Wang, Z.T., Cheng, K.T., Wu, T., Gu, L.H., Hu, Z.B., 2009. Comparison of rDNA ITS sequences and tanshinones between *Salvia miltiorrhiza* populations and *Salvia* species. Bot. Stud. 50, 127-135.

Yang, M., Liu, A., Guan, S., Sun, J., Xu, M., Guo, D., 2006a. Characterization of tanshinones in the roots of *Salvia miltiorrhiza* (Dan-shen) by high-performance liquid chromatography with electrospray ionization tandem mass spectrometry. Rapid Commun Mass Spectrom 20, 1266-1280.

Yang, S., Pang, W., Ash, G., Harper, J., Carling, J., Wenzl, P., Huttner, E., Zong, X., Kilian, A., 2006b. Low level of genetic diversity in cultivated Pigeonpea compared to its wild relatives is revealed by diversity arrays technology. Theor Appl Genet 113, 585-595.

Yin, S.Y., Wang, W.H., Wang, B.X., Aravindaram, K., Hwang, P.I., Wu, H.M., Yang, N.S., 2010. Stimulatory effect of *Echinacea purpurea* extract on the trafficking activity of mouse dendritic cells: revealed by genomic and proteomic analyses. BMC Genomics 11, 612.

Yu, H.C., Kaarlas, M., 2004. Popularity, Diversity, and quality of Echinacea. In: Miller, S.C. (Ed.), Echinacea. CRC Press

Yu, X.Y., Lin, S.G., Zhou, Z.W., Chen, X., Liang, J., Duan, W., Yu, X.Q., Wen, J.Y., Chowbay, B., Li, C.G., Sheu, F.S., Chan, E., Zhou, S.F., 2007. Tanshinone IIB, a primary active constituent from *Salvia miltiorrhza*, exhibits neuro-protective activity in experimentally stroked rats. Neurosci. Lett. 417, 261-265.

Zarzuelo, A., Risco, S., Gamez, M.J., Jimenez, J., Camara, M., Martinez, M.A., 1990. Hypoglycemic action of *Salvia lavandulifolia* Vahl. spp. oxyodon:a contribution to studies on the mechanism of action. Life Sci. 47, 909-915.

Zhang, Y.B., Shaw, P.C., Sze, C.W., Wang, Z.T., Tong, Y., 2007. Molecular authentication of Chinese herbal materials. Journal of Food and Drug Analysis 15, 1-9.

Zhao, Z., Hu, Y., Liang, Z., Yuen, J.P., Jiang, Z., Leung, K.S., 2006. Authentication is fundamental for standardization of Chinese medicines. Planta Med 72, 865-874.

Zhong, G.X., Li, P., Zeng, L.J., Guan, J., Li, D.Q., Li, S.P., 2009. Chemical characteristics of *Salvia miltiorrhiza* (Danshen) collected from different locations in China. J Agric Food Chem 57, 6879-6887.

Zhou, L., Chow, M., Zuo, Z., 2006. Improved quality control method for Danshen products--consideration of both hydrophilic and lipophilic active components. J Pharm Biomed Anal 41, 744-750.

Zhou, L., Zuo, Z., Chow, M.S., 2005. Danshen: an overview of its chemistry, pharmacology, pharmacokinetics, and clinical use. J Clin Pharmacol 45, 1345-1359.

Zhu, S., 2008. Development of a DNA microarray for authentication of ginseng drugs based on 18S rRNA gene sequence. J Agric Food Chem 56, 3953-3959.

APPENDIX 1

Position of the 285 features and 15 controls gridded on each subarray for the *Salvia* SDA

Each subarray was composed of 300 samples. The first 150 samples were printed by one pin and were gridded as follows:

O17	A18	C18	E18	G18	I18	K18	M18	O18	A19	C19	E19	G19	I19	K19
A16	C16	E16	G16	I16	K16	M16	O16	A17	C17	E17	G17	I17	K17	M17
C14	E14	G14	I14	K14	M14	O14	A15	C15	E15	G15	I15	K15	M15	O15
E12	G12	I12	K12	M12	O12/P12	A13	C13	E13	G13	I13	K13	M13	O13	A14
G10	I10	K10	M10	O10	A11	C11	E11	G11	I11	K11	M11	O11	A12	C12
I8	K8	M8	O8	A9	C9	E9	G9	I9	K9	M9	O9	A10	C10	E10
K6	M6	O6	A7	C7	E7	H7	I7	K7	M7	O7	A8	C8	E8	G8
M4	O4	A5	C5	E5	G5	I5	K5	M5	O5	A6	C6	E6	G6	I6
O2	A3	C3	E3	G3	I3	K3	M3	O3	A4	C4	E4	G4	I4	K4
A1	C1	E1	G1	I1	K1	M1	O1	A2	C2	E2	G2	I2	K2	M2

The subsequent 150 samples were printed by a second pin and were gridded as follows:

HPPR	Subtracted sample	DMSO	Ribosomal RNA	Rubisco	a/b binding protein	Cloning vector	Nested F	Nested R	Primer T7	Primer SP6	DMSO 50%	DMSO 50%	Cy5	Cy3
B16	D16	F16	H16	J16	L16	N16	P16	B17	D17	F17	H17	J17	L17	N17
D14	F14	H14	J14	L14	N14	P14	B15	D15	F15	H15	J15	L15	N15	P15
F12	H12	J12	L12	N12	P12/O12	B13	D13	F13	H13	J13	L13	N13	P13	B14
H10	J10	L10	N10	P10	B11	D11	F11	H11	J11	L11	N11	P11	B12	D12
J8	L8	N8	P8	B9	D9	F9	H9	J9	L9	N9	P9	B10	D10	F10
L6	N6	P6	B7	D7	F7	G7	J7	L7	N7	P7	B8	D8	F8	H8
N4	P4	B5	D5	F5	H5	J5	L5	N5	P5	B6	D6	F6	H6	J6
P2	B3	D3	F3	H3	J3	L3	N3	P3	B4	D4	F4	H4	J4	L4
B1	D1	F1	H1	J1	L1	N1	P1	B2	D2	F2	H2	J2	L2	N2

The 15 controls are presented in blue.

HPPR= Salvia miltiorrhiza putative hydroxyphenylpyruvate reductase (hppr) mRNA, partial sequence.

Subtracted sample= aliquot of the enriched *Salvia*-specific sequences obtained from the subtraction process prior to cloning.

DMSO= aliquot of DMSO used to prepare 50% DMSO, which was used to resuspend the PCR products precipitated.

Ribosomal RNA= 5.8S/18S/25S ribosomal RNA sourced from *Cicer arietinum*.

Rubisco= ribulose-1,5-bisphosphate carboxylase/oxygenase gene sourced from *Cicer arietinum*.

a/b binding protein= chlorophyll a/b binding protein gene sourced from *Cicer arietinum.*

Cloning vector= pGEM®-T Easy vector (Promega) digested with *AluI* and *Hae*III an subsequently column purified (QIAquick PCR Purification Kit, Qiagen).

Nested F and R= nested primers 1 and 2R (Clontech) used to PCR amplify the cloned inserts.

Primer T7 and SP6= Primers used to re-amplified the cloned insert from the corresponding isolated plasmid in order to send to sequence.

DMSO 50%= aliquot of the reagent used to resuspend the PCR products precipitated.

Cy5= Dye used as a positive control for the printing process.

Cy3= Dye used as a positive control for the printing process.

APPENDIX 2

Settings for BioRobotics® Total Array System (TAS) Application Suite software v2.6.0.1

The BioRobotics® TAS Application Suite had to be configured before printing. The parameters to print are as follows:

OPTION TAB

Tool Type	Group: 2 Microspot (384 well)
	Tool: 2x1 configuration
Pin refill frequency	
Spots per source visit	
- Salvia	[(8 spots/slide * No. of slides to print) + 20 pre-spots]
- Echinacea	[(10 spots/slide * No. of slides to print) + 20 pre-spots]
Wash frequency	Always wash before pins refills

SOURCE TAB

Microplate options	Microplate group: Generic
	Microplate type: 384 well low profile
	No. of plates: 1
	No. samples: 300
Last plate	150/192, meaning 150 sources on a 384-well plate per pin

Source loading	Hold 1 plate at a time
	Plate have lids
Source action	Dwell

TARGET TAB

EDIT PATTERN

Size	15 x 10 per pin, resulting in a subarray having a 30 x 10
Pitch	0.295 mm (Distance between centers of spots)
Format	Standard

ADAPTER PLATE AND SLIDE LAYOUT

Targets = (No. pre-spoting slides + No. real slides)

Edit layout	Adapter layout 30 vertical slides
	No. of copies fill= targets
Slide layout	Mirror horizontal margins
	X- and Y- spacing adjusted to fit 8grids/slide
	Layout sample set #: 1

TARGET ACTION

Delay before spotting	0.000s
Target height	0.1mm
Dwell time	0.000s

Multiple strikes	1
Pre-spotting:	20 spots
Pitch:	0.700 mm
Slide layout	Mirror horizontal margins
	Top margins 7.50mm
	Bottom margin 8.35mm

EDIT SOFT TOUCH

Soft touch:	Target height 0.1 mm
Soft touch distance:	1.000 mm
Speed:	4.0 mm/s
Climate:	For DMSO buffer,

Target humidity at 60 %

Minimum humidity at 36 %

Bath 1 and 2	Used both baths for 3s
Action	wiggle 0.3mm
Behavior	0.0mm
MWS	Used main wash station for 1 cycle
	Entire wash cycle 2 times

APPENDIX 3

Settings for ScanArray® Express v4.0

The templates used in ScanArray® Express v4.0 were as follows:

Subarrays:

Number of subarrays: 8

Number of columns of subarrays: 1

Rotation (degresss): -0.06

Horizontal pin spacing (mm): 4.5

Vertical pin spacing (mm): 4.5

Spots

Horizontal spot spacing, center to center (µm): 299.398

Vertical spot spacing, center to center (µm): 293.100

Rows of spots per subarray: 10

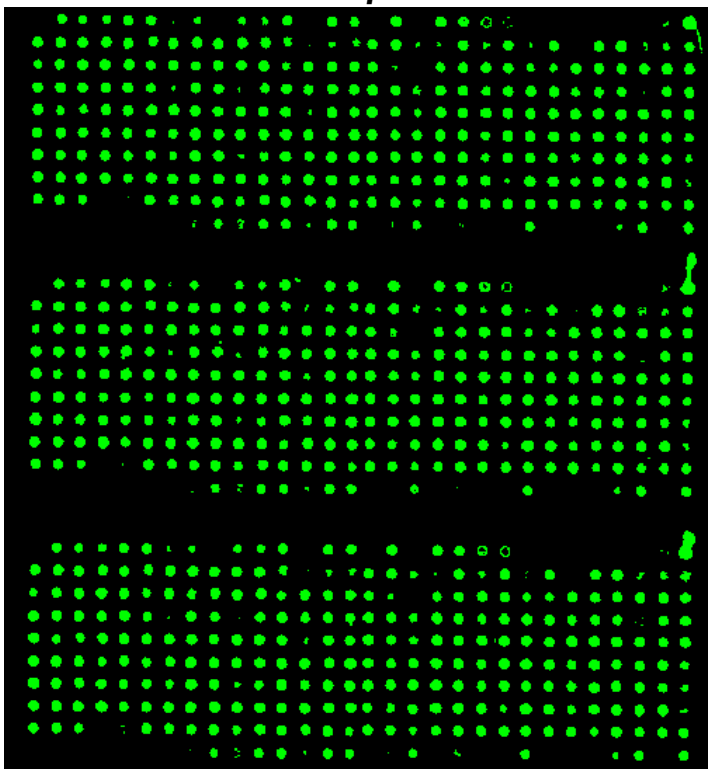
Columns of spots per subarray: 30

Spot diameter (µm): 185

APPENDIX 4

Representative hybridization patterns of *Salvia*

S. miltiorrhiza f. alba



Salvia officinalis

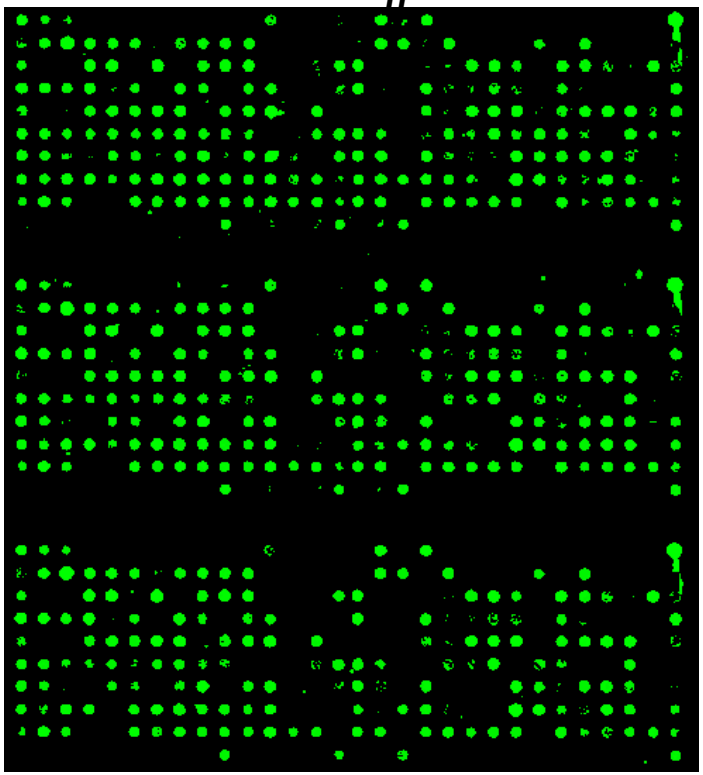


Fig. A4.1. Representative photograph obtained after hybridizing *S. miltiorrhiza f. alba* and *S. officinalis* targets in *Salvia*-SDA

APPENDIX 5

Loading Plots obtained after Principal Component Analysis

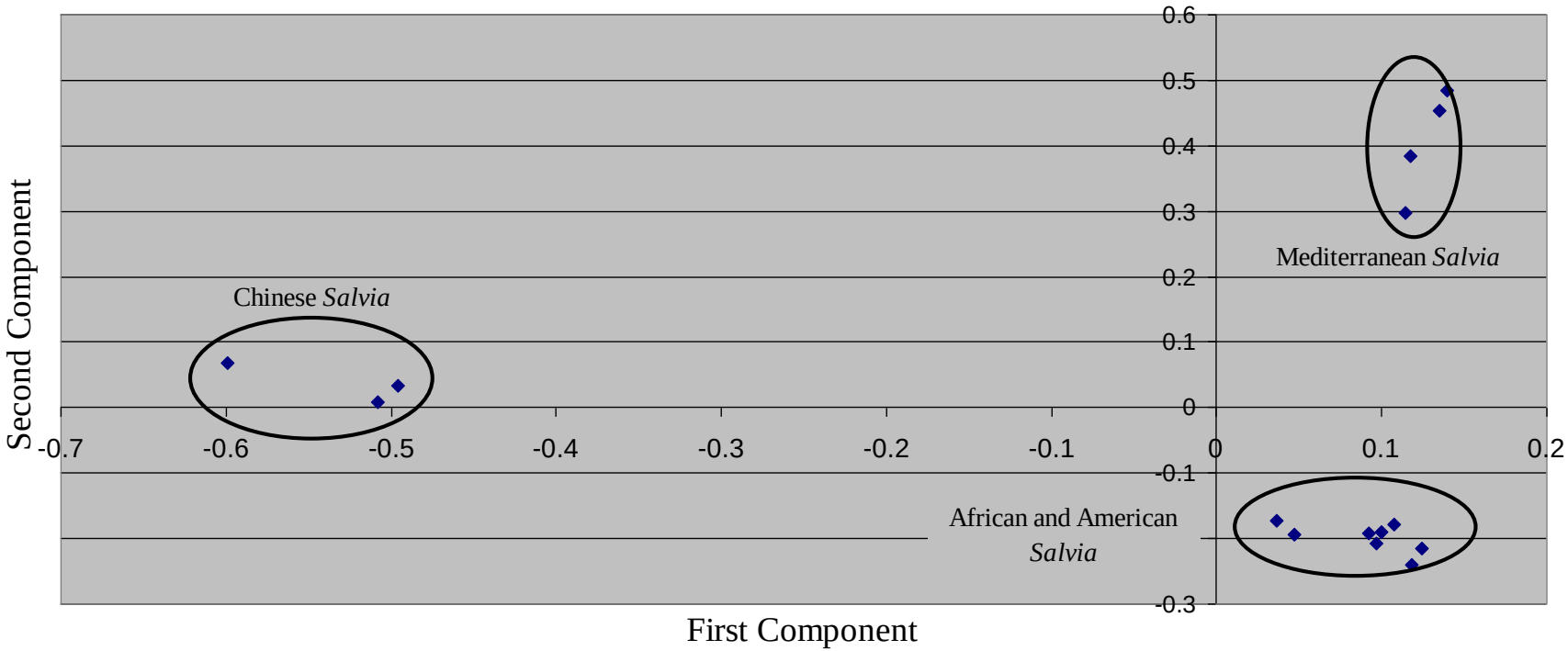


Fig. A5.1. Loading Plots obtained after Principal Component Analysis with the dataset obtained from the SDA hybridization patterns of the fifteen *Salvia* genotypes using the 285 features.

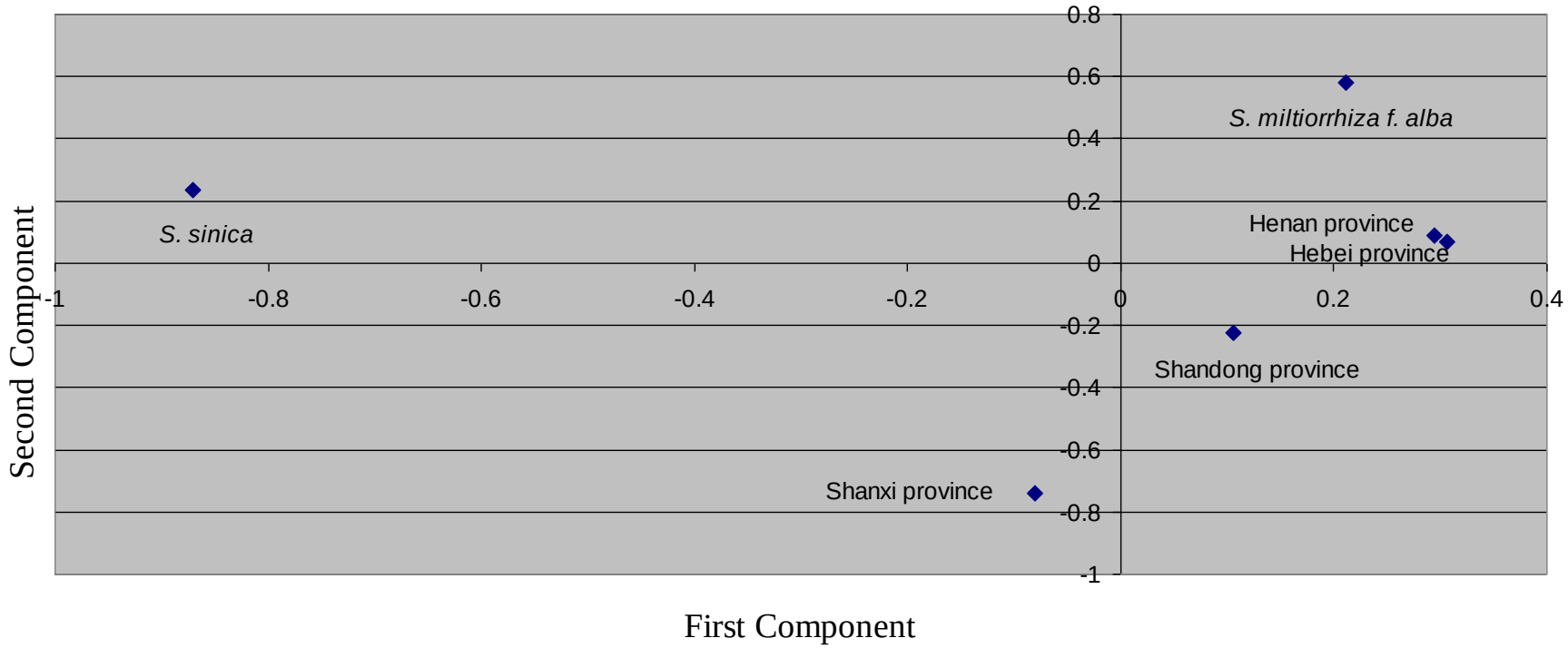


Fig. A5.2. Loading Plots obtained after Principal Component Analysis with the dataset obtained from the SDA hybridization patterns of the five lines of *S. miltiorrhiza* and one of *S. sinica* using 285 features.

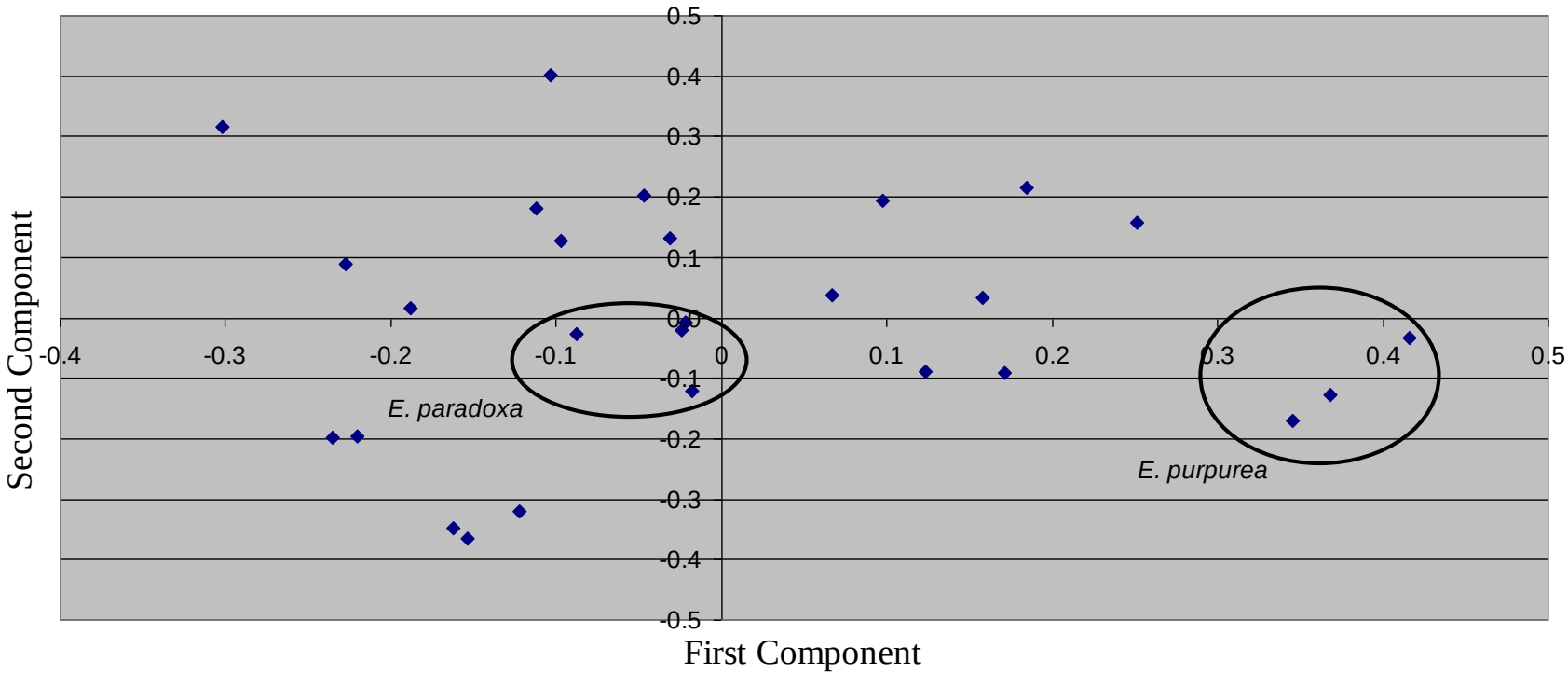


Fig. A5.3. Loading Plots obtained after Principal Component Analysis with the dataset obtained from the SDA hybridization patterns of the 27 Echinacea genotypes using the 283 features.

APPENDIX 6

Pearson bivariate correlation among the highly discriminatory and species-specific features for the fingerprinting of *Salvia* species

Table A4.1. Pearson bivariate correlation of the ten species-specific features and the 4 features chosen by PCA across the fifteen genotypes (SPSS version 17.0). Features with similar patterns of variation were identified. Feature H17 was found to be correlated to features J9 (r = 0.98, P<0.01) and G4 (r =0.99, P<0.01). Also positive significant correlations were found between G13 and N7 (r = 0.99, P<0.01) and between N6 and I7 (r = 0.83, P<0.01).

		N12	H17	J9	G4	E13	O1	F5	G13	N13	N7	N6	I7	P4	A16
N12	Pearson Correlation	1	-.264	-.142	-.219	-.170	-.028	.603*	-.137	-.221	-.155	-.714**	-.722**	-.383	-.200
	Sig. (2-tailed)		.342	.612	.434	.545	.920	.017	.626	.428	.582	.003	.002	.159	.474
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
H17	Pearson Correlation	-.264	1	.978**	.994**	.709**	-.222	-.257	-.134	-.286	-.137	-.301	-.292	-.245	-.606*
	Sig. (2-tailed)	.342		.000	.000	.003	.426	.355	.634	.301	.626	.275	.290	.378	.017
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
J9	Pearson Correlation	-.142	.978**	1	.992**	.719**	-.221	-.168	-.176	-.317	-.182	-.409	-.379	-.333	-.622*
	Sig. (2-tailed)	.612	.000		.000	.003	.429	.550	.531	.250	.515	.130	.164	.225	.013
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
G4	Pearson Correlation	-.219	.994**	.992**	1	.743**	-.212	-.216	-.148	-.304	-.152	-.351	-.328	-.282	-.611*

	Sig. (2-tailed)	.434	.000	.000		.002	.448	.439	.600	.270	.588	.200	.233	.309	.016
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
E13	Pearson Correlation	-.170	.709**	.719**	.743**	1	-.162	-.165	-.149	-.193	-.160	-.348	-.198	-.057	-.364
	Sig. (2-tailed)	.545	.003	.003	.002		.563	.557	.597	.491	.570	.203	.478	.840	.182
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
O1	Pearson Correlation	-.028	-.222	-.221	-.212	-.162	1	-.016	-.041	-.096	-.045	.041	-.083	-.111	-.015
	Sig. (2-tailed)	.920	.426	.429	.448	.563		.954	.884	.733	.873	.885	.768	.694	.957
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
F5	Pearson Correlation	.603*	-.257	-.168	-.216	-.165	-.016	1	-.074	-.150	-.081	-.364	-.369	-.063	.217
	Sig. (2-tailed)	.017	.355	.550	.439	.557	.954		.794	.594	.775	.182	.176	.824	.438
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
G13	Pearson Correlation	-.137	-.134	-.176	-.148	-.149	-.041	-.074	1	.011	.999**	-.059	-.041	.299	.048
	Sig. (2-tailed)	.626	.634	.531	.600	.597	.884	.794		.968	.000	.833	.884	.280	.866
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
N13	Pearson Correlation	-.221	-.286	-.317	-.304	-.193	-.096	-.150	.011	1	.028	.623*	.273	.229	.477
	Sig. (2-tailed)	.428	.301	.250	.270	.491	.733	.594	.968		.922	.013	.324	.411	.072
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
N7	Pearson Correlation	-.155	-.137	-.182	-.152	-.160	-.045	-.081	.999**	.028	1	-.032	-.019	.291	.061
	Sig. (2-tailed)	.582	.626	.515	.588	.570	.873	.775	.000	.922		.910	.948	.293	.829

	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
N6	Pearson Correlation	-.714**	-.301	-.409	-.351	-.348	.041	-.364	-.059	.623*	-.032	1	.834**	.346	.589*
	Sig. (2-tailed)	.003	.275	.130	.200	.203	.885	.182	.833	.013	.910		.000	.207	.021
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
I7	Pearson Correlation	-.722**	-.292	-.379	-.328	-.198	-.083	-.369	-.041	.273	-.019	.834**	1	.307	.567*
	Sig. (2-tailed)	.002	.290	.164	.233	.478	.768	.176	.884	.324	.948	.000		.266	.027
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
P4	Pearson Correlation	-.383	-.245	-.333	-.282	-.057	-.111	-.063	.299	.229	.291	.346	.307	1	.277
	Sig. (2-tailed)	.159	.378	.225	.309	.840	.694	.824	.280	.411	.293	.207	.266		.318
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15
A16	Pearson Correlation	-.200	-.606*	-.622*	-.611*	-.364	-.015	.217	.048	.477	.061	.589*	.567*	.277	1
	Sig. (2-tailed)	.474	.017	.013	.016	.182	.957	.438	.866	.072	.829	.021	.027	.318	
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

APPENDIX 7

Sequences of the of the most discriminatory and species-specific features for the fingerprinting of *Salvia* species

Adaptor 1 is in red and Adaptor 2R is in blue.

>A16. 556bp.

CCCGGGCAGGT
CTCTCAAAGATGTAGATTAGTTACCAAAAGCTCAAACAT
ACAAACTCTCCTATGATAAGTCTAAATACTCAAATATATGAAGACTTAAT
CATATTGTGATCCAAGATTTCCACTAAGTGTTTATTCCCATTAGGACATA
AATCAAATAGCCATTATAAAAATTAAAACCCTTGGATACAAATCTTAGGT
TGAAAGGAAAAGTAGTATTAACAAGACATTATGAAAATTTAACTAAACAT
ACTTACCCTAGAATCAACTGAGAAATTAGTTACTCATATTCAAAGTAAAC
ATAAAGATAGAAATTAAAGACTTGGTAAATAAAAGGGAAGGAATGATAGA
ACCAATAATCTTCAACAAACTTCAATCTTCAACAATGTATCAAGATCCAA
CTATGTAACAAAACTAAAATTTAAAATGTAGAGAGAATGAACTATAAATG
AAAATGTAAGAAAACCAGAATTCTAAAAACTTCTAAGATGTTTCAAGGAA
TGAAAGATTATGAGATATGAGATATTAGATATCCTACAATGCTCAACTAA
GGGGGTATTTATAGGAG **ACCTCGGCCG**

>E13. 373 bp.

GCGGCCGAGGTCCGCGTCA
ATGGGATCTTTTATTCTTGTTGGAGGAGAAATTACCAAACGTCTAGCATT
CCCTCACGCTTGGCGCCAATGAAGTTTTTTTATTTGAGAGAAAAAAGAAA
ACTATGCCTTCGCCATATGAATATTAAGTAATAATAGCATGGCACTTCGA
ATTCGATATGAATTTTTTTTTATTTTTTTTTTCAAAAGATTTGATTATGT
ATCGAGAGAGTAGTATGAGATGAAAGATATTTCCGACTTTCTCTTATCTA
TCGGAAGTCCAATTCAGCGTCACAAACTTATTTGTTTTCACACCGATGGG
CTCTTAACAATATTTAAGTTATAAAAAAAGAGTGCGAAAAAAACCAAATT
TTCTTTTTTGGTTAG **ACCTGCCCGG**

>F5. 218 bp. Same adaptor (Adaptor 2R) at both ends.

GCGGCCGAGGT CTCTAATTAATTAATGGATGTCGGATATTTT
AAATACGGAGATTTAATAAGTCTAAATACAAGCCCCGACTCATCACCGGC
AATAAAGGGGTAAGTCAGTATCGGTTCTCTAGTGGAATGAACTGATATTT
ATAAATTAATTATGGTCTGGGCTGACCATAGATAAATTAATTTATTTGAG
GCCCATCTTTATTCCTTGTATCTGGTCCCTGGACTGG **ACCTCGGCCG**

>G13. 145 bp.

CCCGGGCAGGT CCGGGCAGGTCCG
GTGATTAAAGGCAAGAGGAAAAAACGAATCAATTTCCTCTACCATTCGTA
AATGCGTAAAAGTGGATAGTTTCGTTTTTTGAGGTTGATAGTGTCAAGAA
GGCTCCGCTGAAGAGTAAGATGCTTTCTCTAG **ACCTCGGCCG**

>H17. 612 bp.

GCGGCCGAGGT CTCCTTAAGGGGATGTCATTATCCTATACCGGATA
CGGGTACTAATACAGATAATCAAATATCATATATTAACCGCTATCACCCA
AGATACAGAGTACTCGAGTTAGTATATAACTTTCACCCATAGTAAGTCAA
AGTGATATACGAGTTAACATATATATCTGAATACTTATTAGTATTAAGAT
TTATAAGTCACCGAGATCTTGATTCTTCACTTAAGTCAGATAGAAGAATA
CATCTCAAACTGTGGTCCTATCAATACGTAATGACGTACCAGTATAGACA
AGTAGCCAAGACAAACTACTTCCATCTATACTGTAGCCTAAACCAATAAC
TTGTCCAAGAGTTATTTCGGCTGTGATCATATTATATCTCTTAAGGTTAT
TCCAATTATATGGTCTTCTGTGATCTACAACACACCATATAATCTACTTA
TACAGAGATAAAGAACATACATATGCAATCATGAACACAATCAGATAGGA
GATAAGAATAGTGAAAACATGAATCATTGTATACAAGCATAAAGTTCTTG
CTTTCAGTATACAAATCCAACAATCTCCCACTTATACTAAAGCAAAACTT
TTAGTATACAATGTGTCTAAAAACTAG **ACCTGGCCCGGG**

>N12. 423bp.

CCCGGGCAGGT CTCAATATACTCGATATTTTGGGTAATAGCAATTAT
TATTTGACATGCGATTATATTGCAATAAGGATCCGTGTCCTGCTAATAAC
AGGATGATAATATCCTCTCGAGGAACTTAATAAGTTTATCGTATTAAACC
CTGCAGGTGGAATTAGTTCTGATACGATAATAAGTTTAAGTGGTAGCACT
CGAGATGTCGTTTATAATTAAATAACTAATTAATTAATTAATTGATCGTC
AGAGGAATTAATTAATTAATGGATATTGGATATCTTAAATACGGGGATTA
ATTAAGTCTAATACTAGCCCCGACTTACCTAAAGAATAAAGAGGTAATTC

AGTATTAATTAATTTTCTAGTGGAGTAAATTAATACTTGTGTCTCGATTT
TATTCTGGGCTGAATAAGAAAATCGAGCACTAGGAGG **ACCTCGGCCG**

>N13. 556 bp.

CCCGGGCAGGT CTCTCAAAGATGTAGATTAGTTACCAAAAGCTCAAACA
TACAAACTCTCCTATGATAAGTCTAAATACTCAAATATATGAAGACTTAA
TCATATTGTGATCCAAGATTTCCACTAAGTGTTTATTCCCATTAGGACAT
AAATCAAATAGCCATTATAAAAATTAAAACCCTTGGATACAAATCTTAGG
TTGAAAGGAAAAGTAGTATTAACAAGACATTATGAAAATTTAACTAAACA
TACTTACCCTAGAATCAACTGAGAAATTAGTTACTCATATTCAAAGTAAA
CATAAAGATAGAAATTAAAGACTTGGTAAATAAAAGGGAAGGAATGATAG
AACCAATAATCTTCAACAAACTTCAATCTTCAACAATGTATCAAGATCCA
ACTATGTAACAAAACTAAAATTTAAAATGTAGAGAGAATGAACTATAAAT
GAAAATGTAAGAAAACCAGAATTCTAAAAACTTCTAAGATGTTTCAAGGA
ATGAAAGATTATGAGATATGAGATATTAGATATCCTACAATGCTCAACTA
AGGGGGTATTTATAGGAG **ACCTCGGCCG**

>N6. 250 bp. The other adaptor was not recognized probably due to bad quality of the fragment sent for sequencing .

CTCCCACGCGTCTTGGGGTCCAGGCTCTCCACCCGCGTTGATTCTTATGA
GAAGATGGATCAGAATAATACTATGCTAATCAACACCTTTGGACGTCACA
TTACGTTGACAGGCCCGCGCCGCTGACAATCAATTAAAAAAATATAACTC
AACCCACTTCCCCCTAAATGAACTGAGAAAACCATTACTCATATTCTAAG
TAATCATAACGATTTCAATTTGAGACTTGTAATATAAAAAACTTAGAAAG
ACCTCGGCCG

>O1. 118 bp.

GCGGCCGAGGTCTCTGACA
GCCTTAATTAATTAATCTCTTTTGTAATCCTTAAGCAGTACCACTCAAAC
CTTATTATTGCGTCTGAACTTAATCAACCTGCATGGTTTAGCGCAATAAA
CATTATTGAG **ACCTGCCCGG**

>P4. 526 bp. The other adaptor was not recognized probably due to bad quality of the fragment sent for sequencing .

GCGGCCGAGGT CTACCGCGAAGGCTATCAACTTAGAAA
TGGAGAGCAATTTGAAGAAATGACTTTAAATCTTTGTGTACTGACCCCTA
ATCGAATTGTTTGGGATTCAGAAGTGAAAGAAATCATTTTATCTACAAAT
AGTGGTCAAATTGGCGTATTACCAAATCATGCTCCTATTGCTACAGACCT
GCCCGGGCGGCCGCTCGAATCACTAGTGAATTCGCGGCCGCCTGCAGGTC
GACCATATGGGAGAGCTCCCAACGCGTTGGATGCATAGCTTGAGTATTCT
ATAGTGTCACCTAAATAAAGTCTGAAGGTTTCATTTATATAGGGAGAATT
TTGGAGTGCGCCGCCATAAAATAAAAAAATAGAGGAATAATTCAATCCTG
CGGTAATAAATATAATCTGTTCTTCCTTCCAAAAGAAAGAACTCGGACGA
AACTGTCTTTTTATTTTGATCTGCATGGGCGGCGAATCCTGCTGAATTGC
TGCAATGTGGTTGGGATTATGTCTCCTAGGGAATATGGGTTTAATTAAT

APPENDIX 8

Agronomical traits and the content of four major bioactive constituents of *S. miltiorrhiza*

Table A6.1. Agronomical traits of one-year old plants (Sheng, 2007). Data was calculated from 12 plant-pairs and was shown as a mean.

Line	Number of side branches per plant-pair	Aerial fresh weight (g/plant-pair)	Root number per plant-pair	Maximum root diameter (mm)	Root fresh weight (g/plant-pair)
S. sinica	16 ± 0.6	608 ± 34.1	18 ± 0.6	23 ± 1.2	417 ± 10.2
Shandong province	12 ± 0.3	250 ± 5.7	12 ± 0.3	11 ± 0.3	122 ± 4.4
Shanxi province	12 ± 0.3	253 ± 13.9	13 ± 0.4	21 ± 0.5	315 ± 7.1
S. miltiorrhiza f. alba (**Shandong)**	14 ± 0.8	147 ± 3.9	11 ± 0.3	14 ± 0.3	155 ± 5.3
Hebei province	11 ± 0.3	255 ± 8.9	11 ± 0.3	12 ± 0.3	135 ±6.8
Henan province	15 ± 0.3	328 ± 7.6	11 ± 0.3	12 ± 0.2	214 ± 6.8

Table A6.2. Content of four marker compounds (w/w) [%] in roots of each of the six lines (Sheng, 2007).Data calculated from 3 replicates.

Line	Cryptotanshinone	Tanshinone I	Tanshinone IIA	Salvianolic acid B
S. sinica	0.013 ± 0.002	0.016 ± 0.003	0.069 ± 0.008	4.42 ± 0.26
Shandong province	0.133 ± 0.008	0.043 ± 0.003	0.208 ± 0.013	4.39 ± 0.51
Shanxi province	0.050 ± 0.003	0.029 ± 0.001	0.128 ± 0.008	4.65 ± 0.36
S. miltiorrhiza f. alba (**Shandong)**	0.153 ± 0.007	0.041 ± 0.001	0.207 ± 0.013	3.83 ± 0.17
Hebei province	0.076 ± 0.005	0.024 ± 0.002	0.120 ± 0.009	3.81 ± 0.34
Henan province	0.227 ± 0.016	0.083 ± 0.006	0.351 ± 0.024	4.16 ± 0.33

APPENDIX 9

Pearson bivariate correlation among the highly discriminatory and species-specific features for the fingerprinting of *S. miltiorrhiza* and *S. sinica* populations

Table A7.1. Pearson bivariate correlation of the 10 features detected by PCA (SPSS version 17.0). Features with similar patterns of variation were identified. Feature C11 was found to be correlated to features K6 (r =0.82, P<0.05) and I7 (r =0.85, P<0.05).

		H17	I5	C11	A5	G9	B7	A11	K2	K6	I7
H17	Pearson Correlation	1	-.413	.395	.168	.518	.162	.437	.205	.070	.155
	Sig. (2-tailed)		.415	.439	.751	.292	.759	.386	.697	.895	.770
	N	6	6	6	6	6	6	6	6	6	6
I5	Pearson Correlation	-.413	1	.184	-.574	-.265	-.785	.015	.526	.484	-.080
	Sig. (2-tailed)	.415		.727	.233	.611	.064	.978	.283	.331	.880
	N	6	6	6	6	6	6	6	6	6	6
C11	Pearson Correlation	.395	.184	1	-.282	.259	.015	.655	.756	.822*	.847*
	Sig. (2-tailed)	.439	.727		.589	.620	.977	.158	.082	.045	.033
	N	6	6	6	6	6	6	6	6	6	6
A5	Pearson Correlation	.168	-.574	-.282	1	-.033	.287	-.705	-.475	-.723	.045
	Sig. (2-tailed)	.751	.233	.589		.951	.581	.118	.341	.104	.932
	N	6	6	6	6	6	6	6	6	6	6

G9	Pearson Correlation	.518	-.265	.259	-.033	1	.586	.335	.543	-.046	.285
	Sig. (2-tailed)	.292	.611	.620	.951		.222	.517	.266	.931	.584
	N	6	6	6	6	6	6	6	6	6	6
B7	Pearson Correlation	.162	-.785	.015	.287	.586	1	.136	-.090	-.266	.381
	Sig. (2-tailed)	.759	.064	.977	.581	.222		.798	.865	.610	.456
	N	6	6	6	6	6	6	6	6	6	6
A11	Pearson Correlation	.437	.015	.655	-.705	.335	.136	1	.496	.778	.359
	Sig. (2-tailed)	.386	.978	.158	.118	.517	.798		.317	.068	.485
	N	6	6	6	6	6	6	6	6	6	6
K2	Pearson Correlation	.205	.526	.756	-.475	.543	-.090	.496	1	.659	.573
	Sig. (2-tailed)	.697	.283	.082	.341	.266	.865	.317		.155	.234
	N	6	6	6	6	6	6	6	6	6	6
K6	Pearson Correlation	.070	.484	.822*	-.723	-.046	-.266	.778	.659	1	.540
	Sig. (2-tailed)	.895	.331	.045	.104	.931	.610	.068	.155		.269
	N	6	6	6	6	6	6	6	6	6	6
I7	Pearson Correlation	.155	-.080	.847*	.045	.285	.381	.359	.573	.540	1
	Sig. (2-tailed)	.770	.880	.033	.932	.584	.456	.485	.234	.269	
	N	6	6	6	6	6	6	6	6	6	6

*. Correlation is significant at the 0.05 level (2-tailed). **. Correlation is significant at the 0.01 level (2-tailed).

APPENDIX 10

Sequences of the most discriminatory and species-specific features for the fingerprinting of *S. miltiorrhiza* and *S. sinica* populations

Adaptor 1 is in red and Adaptor 2R is in blue.

> A5. 526 bp.

GCGGCCGAGGT CCGACTTAAGTATTTGATTCTCAAGCCTATCTTTCTACT
GAATATTCACTGACATGTTGGGAACTTACTTGTTCAGTTTTAGAATTTTTG
GTGAAGCCTAAAGTGGTTTTTCCGATTTATGTTCTGAATCTGCCAAATTTG
AACTCTTTTTGTGCACTGAACTTTAGATAGTCATATCTTCCAAACCATGAA
GGTTCTTAGAGCAAATCCAACTGGAGAGTGTTATGATCTCTCCCTAGTTTC
CAGATTGACCTTTGGGGTTCCCAGTGGAGTTTTGTAAAGGGAGATATGAA
TTTTTAAAGAAAGCCCACTGACTTGGTTGTAATCTGGAAATATGAAATTTT
CAGATTTTTGCTTGTTAAATACCCATCTTTGGGAGGGCATATCTTGCTCGTT
TGAATTGTTTTTCATTCGATTCAAATTGGAGAATTATCCTTGAAATGTCTAG
TTTCCAGAATGTCTTTTTGAGCCTCATTTGGAGATCCGAGGTAGATTTGGTG
TCTGTTGTAAAACAAACCCTTAAGAG **ACCTGCCCGG**

> A11. 492 bp. Same adaptor (Adaptor 2R) at both ends.

GCGGCCGAGGT CCATTAGAAACCAAGGCAAACCTTAATAAAAT
TCCATCAAATTTAATCAGCCAAAACTAAAGGTATCCTTGAAGATTTGGTT
AAAATTTCACAAGAGAAAACGTTCATATGATCTGCCAAATAAACAATGAA
ACTCATACACTTTTACTGTTGAAAAAATATGACAGCAGAACAGGGCATTT
TTGAAATAATAAACTGCTCTGTTTCAGAGGTCATAAAAATCGAAATCTTA
TGTTTTTAGAAAAGTATTGAAGTCTAGTTTCGTTTAAAAAAAACGGTGAT
GCAAAATCCTTCATGGATTAATAGATATAAACGTTTTTGTGAAGTCTACC
AACAGTTGACAGATTCTGTCAAGGATTTTTCAAAATATCTTTAAAATAGC
AAACATCATTCAAAAGACATGAAATTTTGCAGCGACGAAATACACATATC
ATAGATAAACATACTAAAATTTCAGAGCCATCAAGCAATGAAAACTCATC
GAAACATAAG **ACCTCGGCCG**

>B7. 218 bp.

GCGGCCGAGGT CCAGTCCAGGGACCAGATACAAGGAATAAAGATGG
GTCTCAAATAAATTAATTTATCTATGGTCAGCCCAGACCATAATTAATTT
ATAAATATCAGTTCATTCCACTAGAGAACCGATACTGACTTACCCCTTTA
TTGCCGGTGATGAGTCGGGGCTTGTATTTAGACTTATTAAATCTCCGTAT
TTAAAATATCCGACATCCATTAATTAATTAGAG **ACCTGCCCGG**

>G9. 341 bp.

GCGGCCGAGGT CTACGTCTCGTCTCATGAAGCCGGATAATTAAC
TAAAATTAATCCGTTCTCTTCGGGGTGTTCTAATCCGTCAATTAAATAAA
CCCCATTCCTCGTAGTCAATAGAGAATAAATAAATACTTCAACTTATTCG
CTTTATAACCTCATAATTAATCGGGCTTAAAAAAATAGGAACAAGTAATG
TTATAAAATTAAATAATTAAAAGGCTCAACATAAGGCAGCCTAATAATTA
ATTAAATAGAATTGGCTTGAATAATTAACATGAGAGGCCCAATTGAAATA
ATTCATCGGCTCAAGTAATTAAATAAATCTTGTCTCAATAAATAAATAAT
TTGATTAG **ACCTGCCCGG**

>H17. 612 bp.

GCGGCCGAGGT CTCCTTAAGGGGATGTCATTATCCTATACCGGATA
CGGGTACTAATACAGATAATCAAATATCATATATTAACCGCTATCACCCA
AGATACAGAGTACTCGAGTTAGTATATAACTTTCACCCATAGTAAGTCAA
AGTGATATACGAGTTAACATATATATCTGAATACTTATTAGTATTAAGAT
TTATAAGTCACCGAGATCTTGATTCTTCACTTAAGTCAGATAGAAGAATA
CATCTCAAACTGTGGTCCTATCAATACGTAATGACGTACCAGTATAGACA
AGTAGCCAAGACAAACTACTTCCATCTATACTGTAGCCTAAACCAATAAC
TTGTCCAAGAGTTATTTCGGCTGTGATCATATTATATCTCTTAAGGTTAT
TCCAATTATATGGTCTTCTGTGATCTACAACACACCATATAATCTACTTA
TACAGAGATAAAGAACATACATATGCAATCATGAACACAATCAGATAGGA
GATAAGAATAGTGAAAACATGAATCATTGTATACAAGCATAAAGTTCTTG
CTTTCAGTATACAAATCCAACAATCTCCCACTTATACTAAAGCAAAACTT
TTAGTATACAATGTGTCTAAAAACTAG **ACCTGGCCCGGG**

>I5. 398 bp.

CCCGGGCAGGT CCCAGCAAAAATTACCTAGTTTTCGAGACC
CCGTTATAAGTCCTATCCATATATGTTCTGATCGCCAACTCATACTTCAT

CCGATTGCATTTTATTGAAATTGAGAGAATACCCCAAATGATTTTGACTT
TACTTTTTTTGTTTCCGAATGAACTTTCATTAACTAGCACTAGTTTGGTG
TGAACTAGCACTAGTTTAATGGGAACTTTTAGGGGTAATTCATCAAATTT
GTTTAGATCTAAAATAATAACATACCCCTCATATGATCCCAATATACATA
TTGATATACATATAGATCGACCAACTTTACATTTTAGGCATCCAGCGATC
CTGATCTATTTGAAACTGGCTAGAATTGAGTTACCTATAGATTAGATCAT
TTTCTGCAGGCATAAGAG **ACCTCGGCCG**

>L5. 407 bp.

GCGGCCGAGGT CTCTCAGTCTGTCAATCCTTACTATGTCTGGAC
CTGGTAAGTTTCCCCGTGTTGAGTCAAATTAAGCCGCAGGCTCCACTCTT
GGTGGTGCCCTTCCGTCAATTCCTTTAAGTTTCAGCCTTGCGACCATACT
CCCCCCGGAACCCAAATACTTTGATTTCTCATAAGGTGCCGGCGGAGTCC
TAAAAGCAACATCCGCCGATCCCTGGTCGGCATCGTTTATGGTTGAGACT
AGGACGGTATCTGATCGTCTTCGAGCCCCCAACTTTCGTTCTTGATTAAT
GAAAACATTCTTGGCAAATGCTTTCGCAGTTGTTCGTCTTTCATAAATCC
AAGAATTTCACCTCTGACTATGAAATACGAATGCCCACGGCTGTCCCTGT
TAATCATTACTCCGATCCCGAAGG **ACCTGCCCGG**

>O14. 504 bp.

CCCGGGCAGGT CTGTTTCTATTTCTACATCTCTTTCTTCCGTTTTTGA
GGTTTCTTTCAGTTTCTTAGTAAAAAAGGGTGACGGTATTCTGCCTAAAT
AGTAGACACAGGTAATAAATAAGAGAATACTAAAGATCCGAGCCATAGAA
TTTATCAATTCTGATACAAGGTACTTATTAGATCGAATGTACTTATTCGA
TCTAATAGAATGATTTTGCCGTATCCAGACTAATACCAATCCAAGCCATT
TCATGAATAAAATGTGACCAATTAACCAACCAACAAAACTACTTGTTACA
AATAACATCTTGTTGTTGCATCGAAACATATAAATGTTGACTAATCTGGC
TAACATTGAACTTGGTAAAATGAAATGGTTGAATAATTGAAAAATGAGAT
TATTCAGGAATAAACATTGAATGCTGAAATTACGCATTGAATTTCTGGTA
GTAGATCCATAATCAAAAAAGTGTTTGTGACTGTTCCAGAAGAAGTGAAA
CAAAAGATATGGTAGAG **ACCTCGGCCG**

APPENDIX 11

Dissimilarity dendrogram for the SDA hybridization patterns of sixteen genotypes including each of the five plants that constitute the pool of Henan and Shanxi province

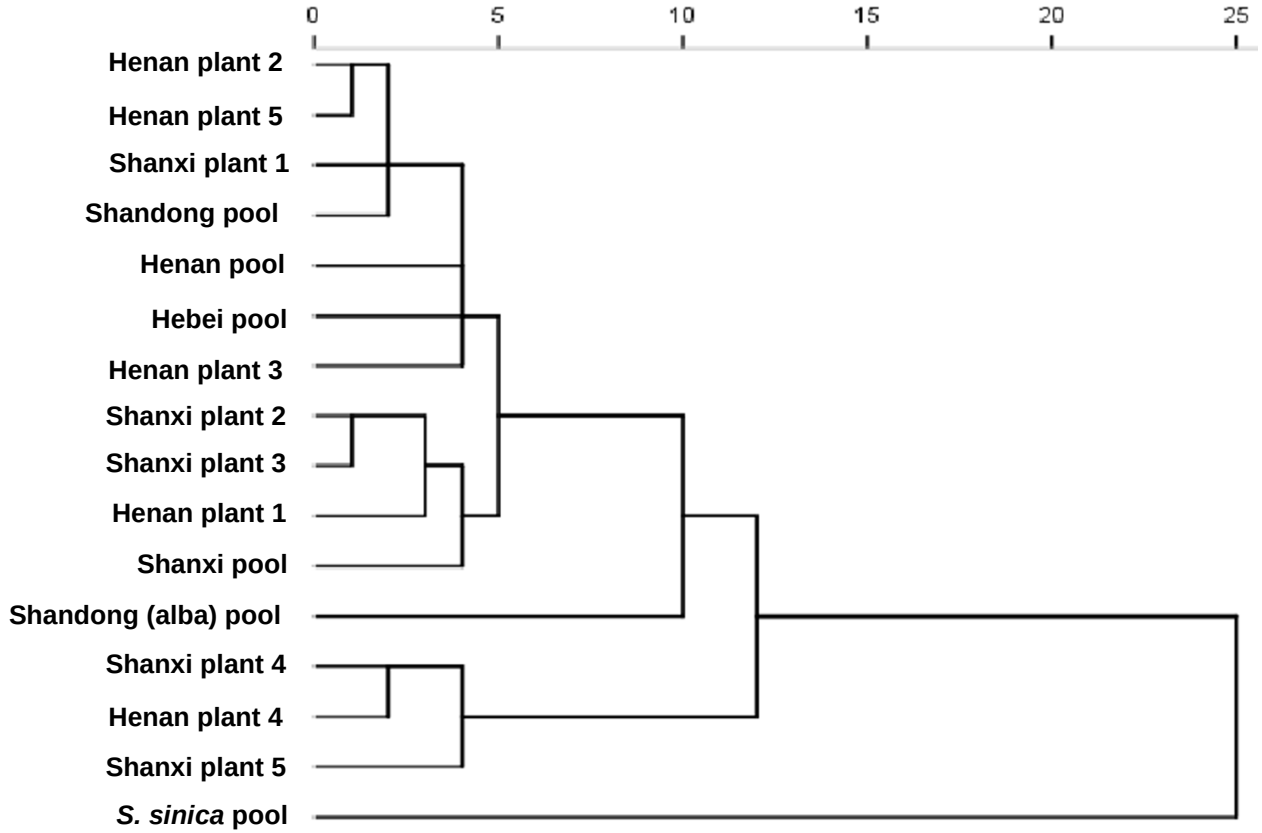


Figure A9.1
Dissimilarity dendrogram (Squared Euclidian distance, between groups linkage) for the SDA hybridization patterns of sixteen genotypes using 285 features. The sixteen genotypes include each of the five plants that conformed the pool of Henan and Shanxi province and the pools of the five lines of *S. miltiorrhiza* and one of *S. sinica*. The steps of the dendrogram show the combined clusters and the values of the distance coefficients at each step; the values have been rescaled to numbers between 0 and 25, preserving the ratio of the distances between the steps.

APPENDIX 12

Position of the 283 features and 17 controls gridded on each subarray for the *Salvia* SDA

Each subarray was composed of 300 samples. The first 150 samples were printed by one pin and were gridded as follows:

O17	A18	C18	E18	G18	I18	K18	M18	O18	A19	C19	E19	G19	DMSO 50%	DMSO 50%
A16	C16	E16	G16	I16	K16	M16	O16	A17	C17	E17	G17	I17	K17	M17
C14	E14	G14	I14	K14	M14	O14	A15	C15	E15	G15	I15	K15	M15	O15
E12	G12	I12	K12	M12	O12	A13	C13	E13	G13	I13	K13	M13	O13	A14
G10	I10	K10	M10	O10	A11	C11	E11	G11	I11	K11	M11	O11	A12	C12
I8	K8	M8	O8	A9	C9	E9	G9	I9	K9	M9	O9	A10	C10	E10
K6	M6	O6	A7	C7	E7	H7	I7	K7	M7	O7	A8	C8	E8	G8
M4	O4	A5	C5	E5	G5	I5	K5	M5	O5	A6	C6	E6	G6	I6
O2	A3	C3	E3	G3	I3	K3	M3	O3	A4	C4	E4	G4	I4	K4
A1	C1	E1	G1	I1	K1	M1	O1	A2	C2	E2	G2	I2	K2	M2

The subsequent 150 samples were printed by a second pin and were gridded as follows:

Rubisco	Cloning vector	Nested F	Nested R	Ribosomal RNA	Rubisco	a/b binding protein	Actin	180 bp cloning vector	180 bp cloning vector	Subtracted sample	DMSO	DMSO	Cy5	Cy3
B16	D16	F16	H16	J16	L16	N16	P16	B17	D17	F17	H17	J17	L17	N17
D14	F14	H14	J14	L14	N14	P14	B15	D15	F15	H15	J15	L15	N15	P15
F12	H12	J12	L12	N12	P12	B13	D13	F13	H13	J13	L13	N13	P13	B14
H10	J10	L10	N10	P10	B11	D11	F11	H11	J11	L11	N11	P11	B12	D12
J8	L8	N8	P8	B9	D9	F9	H9	J9	L9	N9	P9	B10	D10	F10
L6	N6	P6	B7	D7	F7	G7	J7	L7	N7	P7	B8	D8	F8	H8
N4	P4	B5	D5	F5	H5	J5	L5	N5	P5	B6	D6	F6	H6	J6
P2	B3	D3	F3	H3	J3	L3	N3	P3	B4	D4	F4	H4	J4	L4
B1	D1	F1	H1	J1	L1	N1	P1	B2	D2	F2	H2	J2	L2	N2

The 17 controls are presented in blue.

Subtracted sample= aliquot of the enriched *Echinacea*-specific sequences obtained from the subtraction process prior to cloning.

DMSO= aliquot of DMSO used to prepare 50% DMSO, which was used to resuspend the PCR products precipitated.

Ribosomal RNA= 5.8S/18S/25S ribosomal RNA sourced from *Cicer arietinum*.

Rubisco= ribulose-1,5-bisphosphate carboxylase/oxygenase gene sourced from *Cicer arietinum*.

a/b binding protein= chlorophyll a/b binding protein gene sourced from *Cicer arietinum.*

Actin= Actin gene sourced form *Cicer arietinum.*

Cloning vector= pGEM®-T Easy vector (Promega) digested with *AluI* and *Hae*III an subsequently column purified (QIAquick PCR Purification Kit, Qiagen).

Nested F and R= nested primers 1 and 2R (Clontech) used to PCR amplify the cloned inserts.

Partial cloning vector= 180 bp amplified from the pGEM®-T Easy vector (Promega) using primers T7 and SP6.

DMSO 50%= aliquot of the reagent used to resuspend the PCR products precipitated.

Cy5= Dye used as a positive control for the printing process.

Cy3= Dye used as a positive control for the printing process.

APPENDIX 13

Representative hybridization patterns of *Echinacea*

***E. purpurea* (PI631307)** ***E. pallida* (PI631290)**

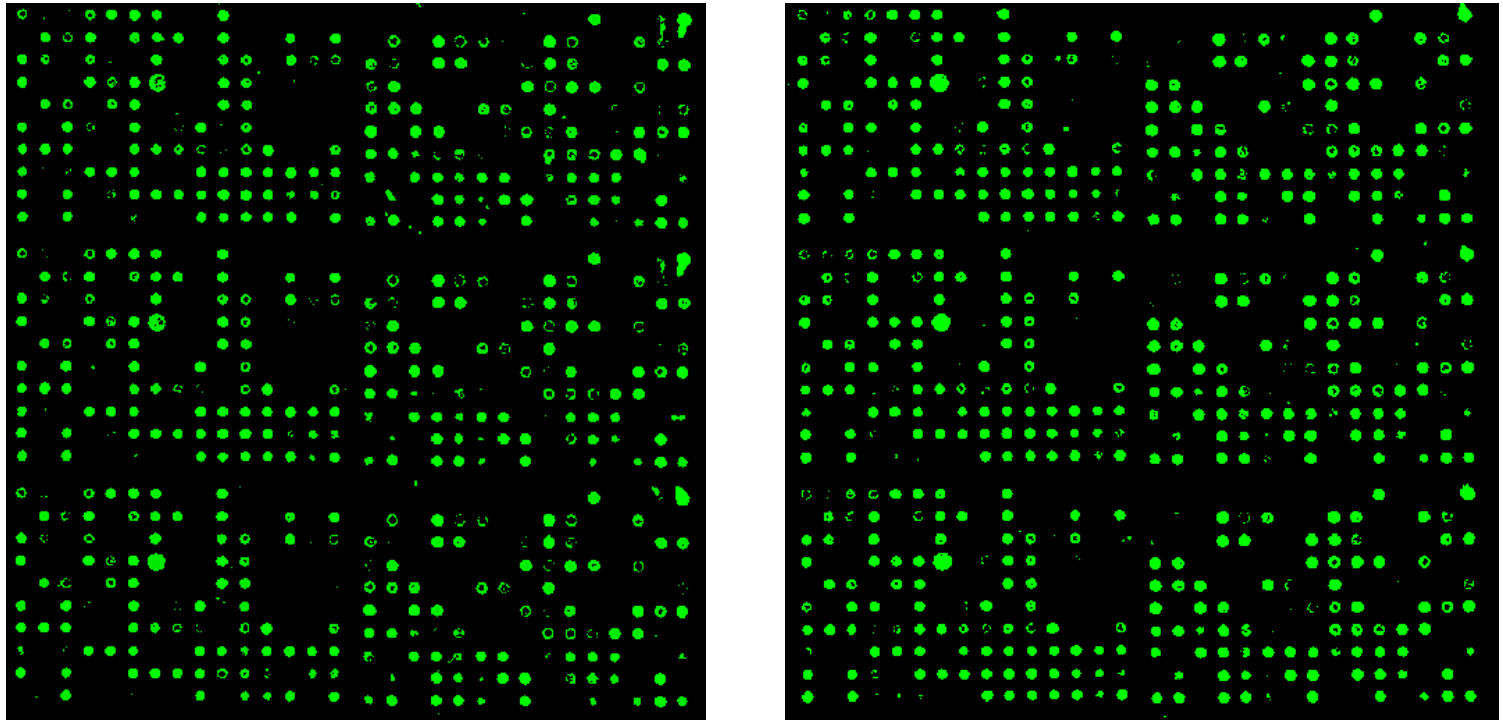


Fig. A13.1. Representative photographs obtained after hybridizing *E. purpurea* and *E. pallida* targets in *Echinacea*-SDA

APPENDIX 14

Pearson bivariate correlation among the highly discriminatory and species-specific features for the fingerprinting of *Echinacea*

Table A11.1. Pearson bivariate correlation of the 11 features detected by PCA (SPSS version 17.0). Features with similar patterns of variation were identified. Feature I9 was found to be correlated to features O2 (r =0.83, P<0.01) and A8 (r =0.84, P<0.01) and feature M2 was found to be correlated to features N6 (r =0.90, P<0.01), A2 (r =0.90, P<0.01) and C2 (r =0.92, P<0.01).

		I9	O2	A8	J8	B15	M2	N6	A2	G16	B17	C2
I9	Pearson Correlation	1	.832**	.846**	.191	-.307	-.537**	-.500**	-.499**	.637**	.091	-.477*
	Sig. (2-tailed)		.000	.000	.340	.119	.004	.008	.008	.000	.653	.012
	N	27	27	27	27	27	27	27	27	27	27	27
O2	Pearson Correlation	.832**	1	.752**	.114	-.579**	-.623**	-.564**	-.556**	.488**	-.075	-.653**
	Sig. (2-tailed)	.000		.000	.572	.002	.001	.002	.003	.010	.708	.000
	N	27	27	27	27	27	27	27	27	27	27	27
A8	Pearson Correlation	.846**	.752**	1	.273	-.156	-.460*	-.368	-.423*	.573**	.172	-.400*
	Sig. (2-tailed)	.000	.000		.168	.438	.016	.059	.028	.002	.390	.039
	N	27	27	27	27	27	27	27	27	27	27	27
J8	Pearson Correlation	.191	.114	.273	1	.214	.169	.232	.180	-.004	.316	.146
	Sig. (2-tailed)	.340	.572	.168		.284	.398	.245	.369	.985	.109	.469
	N	27	27	27	27	27	27	27	27	27	27	27
B15	Pearson Correlation	-.307	-.579**	-.156	.214	1	.772**	.812**	.729**	-.315	.108	.854**
	Sig. (2-tailed)	.119	.002	.438	.284		.000	.000	.000	.110	.594	.000
	N	27	27	27	27	27	27	27	27	27	27	27
M2	Pearson Correlation	-.537**	-.623**	-.460*	.169	.772**	1	.905**	.906**	-.692**	-.155	.918**

	Sig. (2-tailed)	.004	.001	.016	.398	.000		.000	.000	.000	.441	.000
	N	27	27	27	27	27	27	27	27	27	27	27
N6	Pearson Correlation	-.500**	-.564**	-.368	.232	.812**	.905**	1	.823**	-.668**	-.047	.826**
	Sig. (2-tailed)	.008	.002	.059	.245	.000	.000		.000	.000	.818	.000
	N	27	27	27	27	27	27	27	27	27	27	27
A2	Pearson Correlation	-.499**	-.556**	-.423*	.180	.729**	.906**	.823**	1	-.743**	-.297	.937**
	Sig. (2-tailed)	.008	.003	.028	.369	.000	.000	.000		.000	.132	.000
	N	27	27	27	27	27	27	27	27	27	27	27
G16	Pearson Correlation	.637**	.488**	.573**	-.004	-.315	-.692**	-.668**	-.743**	1	.308	-.605**
	Sig. (2-tailed)	.000	.010	.002	.985	.110	.000	.000	.000		.118	.001
	N	27	27	27	27	27	27	27	27	27	27	27
B17	Pearson Correlation	.091	-.075	.172	.316	.108	-.155	-.047	-.297	.308	1	-.207
	Sig. (2-tailed)	.653	.708	.390	.109	.594	.441	.818	.132	.118		.301
	N	27	27	27	27	27	27	27	27	27	27	27
C2	Pearson Correlation	-.477*	-.653**	-.400*	.146	.854**	.918**	.826**	.937**	-.605**	-.207	1
	Sig. (2-tailed)	.012	.000	.039	.469	.000	.000	.000	.000	.001	.301	
	N	27	27	27	27	27	27	27	27	27	27	27

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

APPENDIX 15

Sequences of the of the most discriminatory and species-specific features for the fingerprinting of *Echinacea*

Adaptor 1 is in red and Adaptor 2R is in blue.

>B15. 252 bp.

GCGGCCGAGGT CTACTGAAGTTGGTGGTGGAAGAGTTTG
GGATATGTTTGAAGTTTGTGGCTGTGGAGGGTTAGATTCAAAGATGGGGA
CTGGTATTTGTTGAGGCTGCAATGGTTGTTGTGGGTTTTGTGTTGTTGAT
GGTTCTACGGTATTGGGTTGTACTACCGGGACATCGACATATCCAGGTGA
TTTGGCAGGAGACAACCCTTGTCTAACTTGGGTAGTTAGTATACCATAAG
ATTTGGCTAATCTAGTAATATAAG **ACCTGCCCGG**

>B17. 344 bp.

GCGGCCGAGGT CTAGTCTACCTCTCAGCGCGACTACTTCAA
TACACTTACGAGCAAGGGGTAACTAATAAATCCTCAGAATGAACAGGCAG
CAAAGAATCCTAAGACATCAGGGAACTCTCCTCTCAGAGCCGTTACACCT
AATTAACCTAGCACCTAGACAACCTAACAACTTCACACCTCTCGGCTATC
GAAGATTGGTAGAACGCTAGAGACTTCTAAATCCTTAAAACCCTGTTCCT
ATGGTAATCGTCGACTAAAACCATAAAATAGCTGCTAACAATAAAGTTGC
AACCTAACTAGCCTCATACACGATATAGACAGTCGCAAGAAACTAGTGGA
ACAAGAAATCAAAG **ACCTGCCCGG**

>F15. 744 bp.

CCCGGGCAGGT CCCACAACAAGATGAGATGTGAGAGG
TTTTCAAACAAGTAAAAATTAACTTACCCCTATTGGATGCCATAAACAAA
TTCCAGCATATGCCAAATACTTGAAGGACCTGTGTACCCAAAAACGTCAT
AACAAATTTCCTAAGAAAATTGATTTAACCGAAAATGTCAGTGCCGTTTT
GTCGGATTCCCTTCCTCCTAAACTCCGAGACCCAGGTGCACCTTTGATAC
CCATTCAAGTGGGTGACTTTAAGATGAGCAGGGCACTTTTGGATCTCGGA
GCCAGTGTCAGTATCCTTCCAGGCAGTCTGTACAATCAGTATAATTTCGG
ACCGTTACAGAGAGCCGACACAACTGTTGTGTTGGCTGATTTGAACCTTA
AACTGCCCAGGGGGATTGTCTATGATGTGATTGTCAAAGTTGACAACTTT

TACTACCCAGTTGACTTTTTGGTCTTAGATTATGCTTCAATTGATAAAAA
AAAAACCTAATGTCATACTTGGTAGACCGTTTTTAGCAACCGCGAATGCA
TTAATCGATTGCAGAAACAGAACTGTTGACATCACATTTGGGAATAGGAA
GGTTAGGTTAAATTTGTTTTCCCATACATCTTATCCTCTTGTCAATGATG
AATGCTTCATGGCTGATATCATTGACGGGTGCTCACCTCATACACATGAG
GAAGCCACCATAGAGGCGTGTGCTTTTTTGTGACAGGGAACATTTTGAGC
ATATGGAGGCGTTGGAAG **ACCTCGGCCG**

>G16. 328bp.

CCCGGGCAGGT CTT
AAAGGTATGGGTCATGTATAGGTGAACCTGTTGGAGTGTCAATTGTACTA
CACAGATATATTTACTTTCTTTTAACATTTGTTGAGATTACTCTCTCAAC
GACTTTGTAAACAAATGCTTCCGCTGTGTTAATGAAATGCTAAGTCATAT
TTTGAAAATTGTTCGATATTACTGATGGCTAGATTCTGGGATGTCACGCG
CCTCGCGGTAAAACCCCGCAGTGAAAATTTGAGGGTGTGACAGATTGGTA
TCAGAGCTATTGGTTATAGAGAACTTGGTTTAAAAACTTTTTAAAACCAG
ACTATAACCGTTTTGTTTTCAAAAG **ACCTCGGCCG**

>H9. 249bp.

GCGGCCGAGGT CCACTCCTGTGGGAGTTGGGAATCCTATGGC
GCCATGTGCTGTGGAGGGGATAATGTTGAATTCTGCGATAGCAGGTCTCC
CCAAGATGGTACTATATTGTGAAGTGGAGGTGAGGACGGTGAAAGTTATC
ACCTCAGTTCTGGAGTTGGTTCTATCAGTCAGTGTGACTGGGAACTGAAT
TTCCCCCATTGGGCGGATTATTTCATTGTCGATTCCAGAAATGGTAGTCT
CCACAATCTCCAAGCGAG **ACCTGCCCGG**

>I9. 550 bp.

CCCGGGCAGGT CCGCAAACAATAACACCACCAACACAAT
TTGCCCCGACCGTCCATAACCTATCACTAAGATCAATCCTTGAAAAGGAT
CATCTCAACCATAAAAACTTTATGGATTGGTGTCGAAACTTGAGAATTGT
ACTGAAGTAAGAAAAGAAATACGATGTGTTGGACATCCTATTGCCGATGA
ACCTGACATAGAGGATGAGGATGCTTATCTTGATTAGGTTAGGTATACTG
AGGACTCCGTGTAAGTCTCTTGCCTCATGTTGGCAAGCATGACTCCTGAA
CTCCAAAAGGATTTTGAGAATCACGAGGCATATGACATGATTACCCAGTT
GAAGTAAATGTTCCAACAACAAGCAAGGGTAGAACGCTTCGAAACGGTTC
GAGCACTTCATGCATGTCGTATGGAAGAAATCCAATTGGTCTCATCTTAT

GTCCTCAAGATGAAGAGCCACATTGATCGACTCGAAAGGCTCAATTGTCC
CGTCTCTAAAGAGTTGGCTACGGATTTGATCCTCAACTCTTTAACAAATA
AATTTGAACCATTTGTTATGAG **ACCTCGGCCG**

>I18. 447bp.

CCCGGGCAGGT CTCTTCTA
TGCCGTCGATTGTTCCCTGAGGGAGGTTGTCAAACCAAACCCTCGCTGAC
CCAATGAGCGTCTGGGCAAACATGTGACACCAGACGGGCATTGTCCAACC
TTCTATCCCCCCCGCCGTGTGGAAACCTAATAGATGGTCTTCGGGATCTG
TTGTGCCATCGTATTTGCCGATGGTAGTTGGTATCTTGGTCTTTGGGGGG
AGGGGTGCTCCTGCAATTCTCTTTGTAAATTTTGAGATTTGCAGGAAGGC
TTTTGGTTTATATGGAATAGTAAGATCTTCCGTCCTGTTTTGGCGAGGAA
TATAGGGGTTGTTGAGGTAGACGTTACCTTCAGTATGTTGCGGACCCCTA
TTTTGTGTGGAGTTTGACCCTCTGTCGTTATCAGAGTTGTTTCCATTGTC
TCGTTGGTTTCTGTCCGCAGCTCCGTCTCTTTGGGAAGG **ACCTCGGCCG**

>J8. 300bp. No adaptors found may be due to poor quality at the begging and end of the reading.

CCGCTGGTATTGGTTACAATGGTTTGGTTATTTTGATGGATAATAACTAA
TGGTTTTTCTTAATCAGTAAATTACGGGTTTTCTTAACTAAAACAACTAA
AGTGGTAAATGATAAGATTCAGTGTAACTAATGGCTATGTAGAGATCCTG
GTTTGTCATGCGCCTCGCGGTAAAACCCCGCAGTGGAATTTGAGGGTGTG
ACAGATTGGTATCAGAGCCACTGGTTATAGTGAACTAGCAATGCATCTAG
ACTATAATTGACACTGCATATAAATTAGTTGCACACACTTAGAAATTATT

> L2. 643 bp.

GCGGCCGAGGT CCATTGCAAGTAGTAGATAGTGTGGTGAGT
TTAGCAGATGAGAGTTGGAAGAAGCCATGCGGTGTGGTAAGGGATGTTAT
GATCCAATTGGGTGAATTTCAATACCCGGTGGACTTTTTAGTGTTAGATT
ATGCTTCCACAAATCCATCAGCACAACAAAAAGTTATCTTAGGTCGGCCG
TTTCTACACATGGCCAATGCTCAAATCAATTGTCGGGATGAGGTCATCAC
AATGACCCATGAGAATCGTAAGTTGTTTTTCAATGTTTTGAACAAATCTA
TCACTTATGATGTTGTCACAAAGTTTAATGAATCATGTGTTATTGATGTG
AGTGTTTGTCCACACACTATGGGTGCATGTGAGGGTGTAGGATATGATAA
AGGTATTGGAGATGCAACAGGTGGTAAACCACCGGATTGTGATGAAGAAC

ATATGGGGTCAAGTAGGATTGTCTTGGATTACCTACCGAATGGACATGTT
TTGGATGCATGTAGTGTTGGAGATGGCGGGAACCGCTTCTTTGAGCCACC
ATAAATGGGGGTGGCACGGTCTGGCTGAAGACCCGAAAACTTAGCGCTGC
TCGGGAGGCAACCCGAGGCTTTACACTAATATGTTCAGTTTCTTTACCTT
TAGTGTTTCAGGG **ACCTGCCCGG**

>M2. 829 bp.

GCGGCCGAGGT CTAAGAT
CCTATGTCAGGAGAAGGTAGTTAGCATTCCTCTTCCTGATGGGGAAACTC
TTGTAATCCAAGGAGAACGAAAAGATGTACCTATAAACATCATCTCCTCA
ATGAAAGCCCAGAAATGGATACGTAAGGGCTATCCTGCCATTATGGCACT
CGTGACCGATACTCGCACTGAAGAGAGTAAACTTGAAGATTTCCCAGTTG
TGCGAGATTTTCCTGATGTATTCCCAGAGGATTTGCCTGGATTACCCCCA
CACCGACAGGTCGAATTCCAGATTGACCTTACTCCAGGTGCAGCACCAGT
AGCCCGCGCACCATACAGATTAGCACCAGCTGAACTACTGGAACTGTCAA
CTCAACTTCAAGAACTGTTGGACAAGGGTTTTATACGTCTTAGTTCATCT
CCATGGGGAGCACCAGTCTTGTTCGTCAAGAAGAAGGACGGAACCTTCCG
CATGTGTATCGATTACCGGGAACTAAACAAAGTGACAATCAAAAATCGTT
ATCCACTACCTCGTATCGATGATCTCTTTGACCAACTGCAAGGGTCGAGT
TATTATTCGAAGATCGATCTGCGATCTGGTTATCACCAGTTGAGGGTACG
TGAGGAGGACGTATCCAAGACTGCTTTTCGTACTCGGTATGGCCACTTCG
AGTTTCGGTCATGCTATTCGGCCTTACAAACGCACCAGCGGTATTCATAG
ACCTCATGAATCGTGTGTGCAAACTGTATCTGGACAAGTTCGTCATTGTG
TTCATAGATGACATCTTGATCTACTCCAAGAGCAAGGATGAGCACGCGGA
GCACTTGCGTCTAATCCTGGAG **ACCTGCCCGG**

>M8. 454 bp.
GCGGCCGAGGCTCACAGTTTTAAGACTCCAAAAACTGTTGTAACCTTTCTT
TTCAAAACATAGACAGTTTGTATAGTATTATTATGATTATATTCAGATCAAG
TTATTACTTTGCAAAGGCAATAACAATTGTGCATTCACATAAGACCATAATT
AAACCAAACAATCCATGAACCGGATTAGGTCACAATTGCATACTCCCGAGG
CTGGACCTAATCTGCCAGTTCAATCCTCCGAATATGGGGCTTGTTAAACCCG
ATAGATCTATCCAACAGTACCGAGGTCAATGATTAATAATTATGTACCGTTT
ATATGTCCACGGTGTGCTCCAATCTCATGCACACAATCAGATCAATATTATC
AAATAATTCTACAAATAGTTTCACATACATGTATCTCCCCCATAGTTTAAAA
CATATCAAAACAGTTAAAAAGGGGCTGCGAAACTCACTGTCGTGTAATAG
ACCTGCCCGG